САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА ТЕХНОЛОГИИ ПРОГРАММИРОВАНИЯ

Синельникова Анастасия Сергеевна

Магистерская диссертация

Исследование существующих систем моделирования деятельности головного мозга с целью выявления проблемных участков и несовершенств ПО высокой сложности

Направление 02.04.02

Фундаментальная информатика и информационные технологии

Магистерская программа технологии баз данных

Научный руководитель,

доктор физ.-мат. наук,

профессор

Богданов А.В.

Санкт-Петербург

2018

Содержание

Введение

Проблема исследования мозга человека, проблема соотношения мозга и психики – одни из самых сложных задач, которые ставились в науке. Ученые поставили цель познать нечто, равное по сложности самому инструменту познания. Не приборы и не методы – главное средство познания, им остается – мозг человека.

К решению сверхсложной задачи начали подходить с исследования мозга мыши. Мозг мыши дает представление о базовых строительных блоках, которые относятся ко всем млекопитающим, в том числе к людям. Вычислительную мощность анализа мозга мыши осуществляет Blue Gene/Q – мощнейший суперкомпьютер фирмы IBM.

На сегодняшний день с помощью Blue Gene/Q исследователи из IBM смоделировали только 4,5% нейронов и синапсов мозга человека. Исследователи постепенно изучают работу мозга, разбивая его на маленькие части, что говорит о том, что все функции и возможности мозга не рассматриваются. Кроме того, составляя примерную оценку вычислительной мощности суперкомпьютеров, которые смогли бы обрабатывать всю информацию мозга, потребуется ни один компьютер Blue Gene, а тысячи таких компьютеров, которые территориально займут небольшой городок. Производством энергии для этих суперкомпьютеров занималась бы целая атомная станция, но при этом наш мозг обходится малым потреблением энергии, например, из пищи. Даже, если в ближайшее время ученым удастся создать физическую модель мозга, допустим, кошки и выстроить его функционирование и деятельность, кошка с таким мозгом не сможет поймать мышь. В таком мозге нет теменной доли, поэтому нет сенсорных и моторных связей с внешним миром, базовые связи не представляют мыслительные процессы кошки. В таком мозгу нет ни обратных связей, ни воспоминаний о выслеживании добычи или поиске партнера. Компьютеризированный мозг

кошки — чистый лист бумаги, лишенный всяких воспоминаний и инстинктов.

Одна из целей исследования мозга — возможность понять механизмы возникновения различных неврологических заболеваний и деменции. Миллиарды людей страдают от психических заболеваний, причина которых до сих пор неизвестна, неясно, что именно в мозге работает не так — какой путь, синапс или нейрон. Возможно, патологии вызваны не массовым разрушением нейронов, а их неверной коммутацией.

Одним из примеров такого нарушения может стать синдром Капгра, при котором вы видите женщину, узнаете в ней свою мать, но считаете, что эта женщина — самозванка. По мнению индийского невролога, психолога и доктора медицины В.С. Рамачандрана эта редкая болезнь может вызываться нарушением связи между двумя частями мозга. За узнавание лица матери отвечает веретенообразная извилина височной доли мозга, а за эмоциональную реакцию на это — мозжечковая миндалина. Если связь между этими отделами мозга нарушена, человек узнает свою мать, но, поскольку эмоциональной реакции на это не возникает, он убеждается, что это самозванка [1]. Кроме того, схема мозга помогла бы разрешить такие базовые, но до сих пор не проясненные вопросы, как принцип хранения долговременных воспоминаний. Известно, что определенные части мозга, такие как гиппокамп и мозжечковая миндалина, отвечают за хранение воспоминаний, но до сих пор не ясно, как эти воспоминания распределяются по разным участкам коры, а затем собираются воедино [2].

Актуальность. Исследование мозга является важнейшей задачей человечества не только для понимания биологических механизмов, наших мыслей, эмоций, сознания, – которые и делают нас людьми, но и по практическим соображениям. Понимание мыслительных процессов мозга и сознания человека даст возможность создавать кардинально другие, мощнейшие вычислительные системы, ставя под сомнение привычное технологическое мироустройство. Исследование мозга так же необходимо

для понимания, диагностики и лечения заболеваний, что является глобальной проблемой в современном мире. Ответы на вопросы о деятельности мозга исчерпают потребность проводить эксперименты на животных, что приведет человечество к новому культурному развитию.

В мировом сообществе существуют несколько ведущих проектов, занимающихся разработкой программного обеспечения по моделированию мозга: Blue Brain Project, Human Brain Project, Human Connectome Project, Allen Human Brain Atlas, Watson (IBM), DARPA SyNAPSE. Проекты направлены, в основном, на использование нейроморфных вычислений – применяя искусственные нейронные сети в комбинации со специализированными чипами, архитектура которых напоминает структуру мозга и является аппаратной поддержкой нейронных сетей, а также на реализацию механизмов нейро-физико-химического взаимодействия нейронов между собой и окружающей средой. Это основа, которая может дать ответы на фундаментальные вопросы о работе самих нейронов, – передачи информации через нервные импульсы от нейрона к нейрону благодаря различным конфигурациям нейронов и синапсов. Однако, так как человеческий мозг содержит 86 миллиардов нейронов, каждый из которых имеет в среднем 1750 связей с другими нейронами, текущей мощности суперкомпьютеров недостаточно для моделирования всего человеческого мозга [3].

Постановка задачи

Проблемой проектов по моделированию мозга человека является отсутствие ориентированности на важный аспект, благодаря которому человек часто изменяет принятые решения, – чувства и эмоции, психика и сознание, которые выделяют человека среди других биологических видов.

Отличительной особенностью данной работы является тот факт, что в России на сегодняшний день нет масштабных проектов, связанных с моделированием человеческого мозга и мышления. Это является брешью во многих сферах государственного устройства, ибо это не только проблема и загадка науки, но и политический, экономический вопрос. Санкт-Петербургский государственный университет совместно с Национальным медицинским исследовательским центром психиатрии и неврологии имени В.М. Бехтерева видит перед собой цель в создании экспериментальной установки, напоминающей человеческое мышление, которую можно будет использовать не только в медицине (например, диагностика деменции на ранних стадиях), но и в совершенно разных областях: финансы, промышленность, анализ данных, маркетинг, энергетика и др.

Объектом данного исследования является оптимальная конфигурация экспериментальной установки для моделирования мозга человека, на основе альтернативного подхода, учитывающего функциональную асимметрию мозга и человеческого сознания.

Цели и задачи. Целью данной работы является исследование существующих систем моделирования деятельности головного мозга для выявления проблемных участков и несовершенств программного обеспечения (ПО) высокой сложности.

Для достижения поставленной цели нужно решить следующие задачи:

1) анализ существующих решений среди проектов по моделированию мозга;

2) анализ архитектур TrueNorth и Power9, используемых для моделирования деятельности мозга в задачах искусственного интеллекта;

3) исследование особенностей функциональной асимметрии головного мозга человека;

4) тестирование компонентов экспериментальной установки;

5) создание оптимальной конфигурации экспериментальной установки для задач моделирования деятельности мозга с применением альтернативного подхода, учитывающего асимметрию головного мозга.

Методология и методы исследования. Практическая составляющая данной работы заключается в тестировании задержки и пропускной способности сетей Ethernet и InfiniBand с помощью Intel MPI Banchamark, а так же тестировании производительности вычислительного кластера с двумя картами GPU NVIDIA Tesla P100 с помощью тестов High Performance LINPACK (подробнее о методах и результатах пойдет в соответствующих главах).

Тестирование проводилось на виртуальной машине Вычислительного центра СПбГУ.

Обзор литературы

DARPA SyNAPSE Program

SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) – это программа, разрабатывающая нейроморфные микропроцессорные системы на базе не фон-Неймановской параллельно распределенной масштабируемой архитектуры, которые построены по принципам работы мозга животных и требующие низкого энергопотребления. Основным достижением программы является самый большой на сегодняшний день цифровой нейроморфный чип TrueNorth, о котором пойдет речь в разделе 2.1 [4].

Blue Brain Project

Проект Blue Brain Project (BBP) стартовал в 2005 году совместно с IBM и EPFL (Швейцарский Федеральный Технический Институт Лозанны). Основная цель проекта – создание биологически подробных цифровых реконструкций и моделирование мозга мыши, а в дальнейшем мозга человека. Моделирование сознания задачей исследования не является [5].

Одно из важных достижений BBP было создание модели колонки неокортекса. Неокортекс – часть коры головного мозга, отличие мозга человека от мозга животных заключается в степени развития данной коры. Неокортекс отвечает за сенсорное восприятие, моторику, пространственную ориентацию, мышление и речь. Нейроны в неокортексе формируют вертикальные колонки – фрагменты диаметром около 0,5 мм и высотой около 2 мм. У человека неокортекс содержит примерно 500 тыс. колонок, каждая из которых состоит около 60 тыс. нейронов.

В 2007 году была создана модель колонки неокортекса крысы (10 тыс. нейронов) и были представлены основные достижения проекта на тот период: была создана новая модель сеточной структуры, которая

автоматически генерирует нейронную сеть по биологическим данным, новый процесс симуляции и саморегуляции, который автоматически проводит проверку модели для более точного соответствия биологической природе, первая модель колонки новой коры клеточного уровня, построенная именно по биологическим данным.

В 2015 г. была создана детальная реконструкция небольшого участка соматосенсорной коры мозга мыши, состоящей из 31 тыс. нейронов, связанных друг с другом 8 млн. связей. Соматосенсорная кора – это область, которая получает осязаемую информацию, когда усы мыши и другие части ее тела прикасаются к чему-то [6]. На данный момент симуляция этого небольшого участка мозга требует от суперкомпьютера около миллиарда вычислений каждые 25 микросекунд, а симуляция мозга человека потребует больше вычислений в миллиард раз [7].

В данный момент основными задачами проекта являются:

1. Создание модели на молекулярном уровне, для изучения эффекта экспрессии генов (процесса, в котором наследственная информация от гена (последовательность нуклеотидов ДНК) преобразуется в РНК или белок. На рисунке 1 представлена имитированная сеть нейронов.

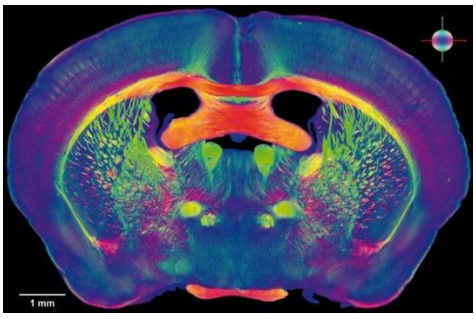

Рис. 1. Имитированная сеть нейронов, показывающая трехмерные формы отдельных клеток (реконструированные из лабораторных данных)

2. Упрощение модели колонки для параллельной симуляции большого количества соединённых колонок для создания модели неокортекса человека.

Концепция Builder. Вся цифровая реконструкция мозговой ткани BBP создается с помощью программного приложения Builder, которое генерирует компьютерную модель конкретной структуры мозга. Цифровые реконструкции основаны на экспериментальных данных и на математических абстракциях. Стратегия Blue Brain требует построения реконструкций, представляющих различные уровни организации мозга. Каждый из них требует определенного программного компонента:

1. Cell Builder – для построения модели отдельных нервных клеток. На рисунке 2 представлены связи нервных клеток.

2. Mesocircuit Builder – моделирование нейронных схем, охватывающих несколько колонок неокортекса, модулей или микросхем.

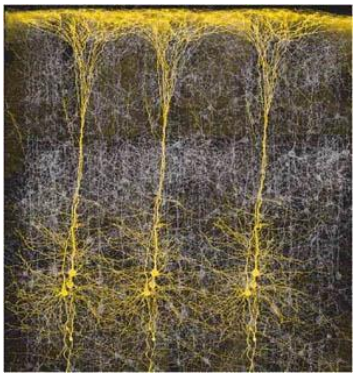

Рис. 2. Связи нервных клеток

3. Experiment Builder – создание явных описаний экспериментальных сред и протоколов для экспериментов [8].

На рисунке 3 представлен пример программной среды для моделирования и анализа экспериментальных данных, содержащий несколько разнонаправленных компонентов.

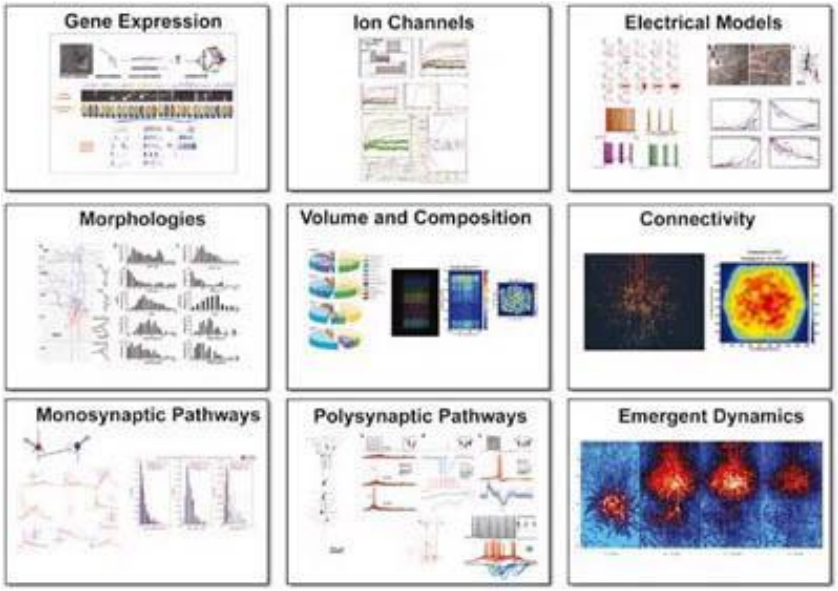

Рис. 3. Программная среда для моделирования и анализа экспериментальных данных [2]

Иными словами, данный проект пытается смоделировать анатомическую, биологическую модель мозга. Фундаментально «понять» действие нейронов, как синапсы взаимодействуют между собой и т.д.

Human Brain Project

Human Brain Project (HBP) создает исследовательскую инфраструктуру, помогающую продвигать нейронауку, медицину и компьютерные вычисления. Глобальная цель HBP – понимание работы головного мозга человека. Это один из крупнейших научных проектов, когда-либо финансируемых Европейским Союзом [9].

10-летний проект начался в 2013 году и в нем непосредственно работают около 500 ученых в более чем 100 университетах, учебных больницах и исследовательских центрах по всей Европе [9].

Шесть исследований платформы составляют основу инфраструктуры HBP:

1. нейроинформатика (доступ к общим данным о мозге);

2. мозговое моделирование (репликация архитектуры мозга и активности на компьютерах);

3. высокопроизводительная аналитика и вычислительная техника (обеспечивающая требуемые вычислительные и аналитические возможности);

4. медицинская информатика (доступ к данным пациентов, идентификация сигнатур болезни);

5. нейроморфные вычисления (разработка мозговых вычислений);

6. нейроробототехника (neurobotics – использование роботов для тестирования мозговых моделей).

Все эти направления связаны между собой и включены в единую открытую платформу Brain Simulation Platform (BSP) для экспериментов с симуляцией функций человеческого мозга, ученые со всего мира могут объединяться, сравнивать свои результаты и предлагать свои решения [9].

Платформа BSP включает набор программных инструментов и рабочих процессов для совместного исследования мозга, чтобы позволить исследователям реконструировать и моделировать детализированные многоуровневые модели мозга, отображая возникающие структуры и поведение. Исследователи могут выбрать необходимый уровень детализации в соответствии с научными вопросами [10].

С помощью платформы были получены три основных достижения:

1) большая валидация стратегии реконструкции и моделирования, лежащей в основе BSP;

2) выпуск BSP в интернет-доступе для научного сообщества и общественности;

3) первое создание целых пирамидных моделей человеческого нейрона.

На рисунке 4 представлены модели мозга при МРТ и КТ-визуализации.

Рис. 4. МРТ и КТ-визуализация для реконструкции подкорковых и корковых поверхностей

Human Brain Project помимо создания атласа мозга и его моделирования старается разобраться и в других важных составляющих деятельности мозга: неизведанное сознание, память и процесс познания человеком окружающего мира [10].

Представьте себе яблоко – его зелень, кислый вкус и свежий хрустящий хруст; как мозг создает представление о таком яблоке? Это один из вопросов, задаваемых HBP. Вопрос имеет решающее значение, поскольку эти представления являются основой для более высоких когнитивных процессов, таких как формирование категорий, рассуждение и язык. Одна из целей проекта заключается в разработке сети нейронных систем с глубоким обучением, которая учится распознавать объекты и функции так же, как и реальные нейробиологические системы [9].

HBP поддерживает работу когнитивных и теоретических нейробиологов, чтобы развить более глубокое понимание работы нашего мозга. Теоретические нейробиотики работают над разработкой многомасштабной теории мозга, которая синтезирует нисходящие и основанные на данных восходящие подходы. Они также пытаются: объединить теории обучения, памяти, внимания и целенаправленного поведения; понимать сложные когнитивные функции, такие как пространственная навигация, рекурсия и символическая обработка; и идентифицировать мосты, связывающие множественные временные и пространственные масштабы, связанные с деятельностью мозга [9].

Когнитивные нейробиологи HBP изучают природу зрительного восприятия, распространение медленных волн в спящем мозге, роль гиппокампа в эпизодической памяти (личные воспоминания из нашей жизни), а также разработку новых способов измерения состояния сознания [9].

Эпизодические воспоминания – это личные, осознанные переживания, установленные в пространстве и времени. Способность мозга вспоминать объекты и опыт из мультисенсорной информации (например, видение, слух или прикосновение) является ключом к пониманию человеческой памяти. HBP проводит скоординированную серию экспериментов для идентификации нейронных механизмов за эпизодической памятью и проверяет их с помощью вычислительных моделей и роботизированных систем [9, 10].

Ученым нужна поддержка огромных вычислительных мощностей, на данный момент человечество не способно воспроизвести вычисления равные вычислениям мозга.

HBP имеет распределённую сеть суперкомпьютеров с центрами в Германии, Италии, Испании и Швейцарии. В арсенале одни из самых мощных компьютеров в мире, способные выполнять квадриллионы операций в секунду с объемом памяти равным квадриллионам байтов. Один из суперкомпьютеров настолько же мощный, как и около 350 000 стандартных компьютеров.

Данный проект, как было сказано, имеет большое финансирование, благодаря которому может оперировать бюджетом в удовлетворении любых потребностей, касаемо увеличения мощности систем, повышения квалификации сотрудников, разработки уникального программного обеспечения и т.д. К сожалению, в рамках данного исследования мы не можем идти по пути описанного проекта.

Allen Human Brain Atlas

Институт Аллена наук о мозге основан в 2003 году меценатом Полом Алленом для изучения работы мозга. Цель проекта — составить карту или атлас мозга с упором на гены, ответственные за его формирование. Составление карты мозга мыши завершилось в 2006 году, следующий шаг – создание анатомически и генетически полной трехмерной карты человеческого мозга. Можно надеяться, что понимание того, как в мозге происходит экспрессия генов, поможет разобраться в механизмах аутизма, болезни Паркинсона, синдрома Альцгеймера и других расстройств [11].

Проект является общедоступным через портал Allen Brain Atlas на сайте brain-map.org., можно наблюдать за развитием проекта или принять в нём участие. Каждый ресурс Allen Brain Atlas объединяет данные из тысяч экспериментов, давая беспрецедентные трехмерные справочные пространства для генетической и анатомической информации о человеческом и мышином мозгах, это мощный способ изучения данных экспрессии генов, нейронных связей, клеточных характеристик и нейроанатомии.

Проект имеет множество публичных программ по исследованию мозга человека, мыши, спинного мозга и др. Рассмотрим некоторые из них.

Атлас человеческого мозга – это уникальный мультимодальный атлас, который отображает экспрессию генов через мозг человека. Интегрируется анатомическая и геномная информация посредством магнитно-резонансной томографии (МРТ/MRI), диффузионной тензорной томографии (ДТТ/DTI), гистологии и данных экспрессии генов, полученные как из микрочипов, так и подходов гибридизации. Возможен интерактивный просмотр с помощью программного обеспечения Brain Explorer 3D на портале Allen [11].

На рисунке 5 представлен атлас Brainnetome человеческого мозга [12].

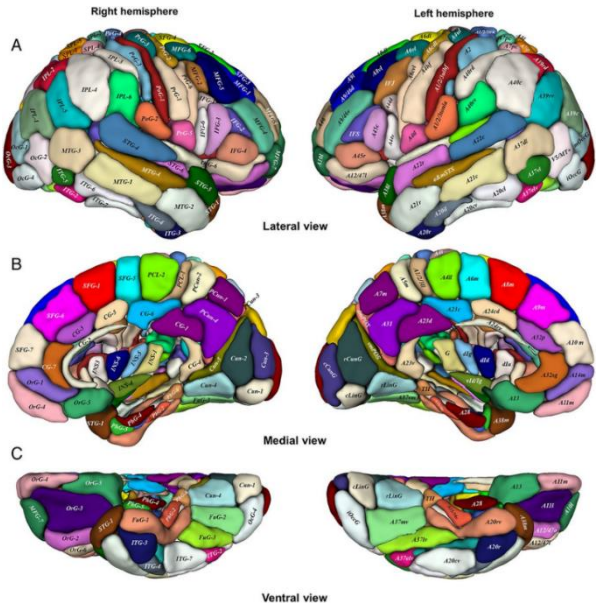

Рис. 5. Схема разделения на части человеческого мозга в Атласе Brainnetome

BrainSpan – атлас развития человеческого мозга (Atlas Of The Developing Human Brain). Атлас обеспечивает широкий и подробный анатомический анализ экспрессии генов в развитии человеческого мозга, включающий гибридизацию, секвенирование РНК и подходы к микрочипам наряду с поддерживающим нейроанатомическим эталонным контентом. Атлас был разработан консорциумом научных партнеров из нескольких организаций и финансировался наградами Национального института психического здоровья. Этот ресурс напрямую доступен на сайте brainspan.org.

Исследование старения, слабоумия и травматического повреждения мозга (Aging, dementia, tbi study). Проект по проблемам старения, слабоумия и травматического поражения головного мозга представляет собой подробную невропатологическую, молекулярную и транскрипномическую характеристику головного мозга.

Проект предоставляет исследователям API с доступом к обширным наборам данных Института Allen: **Allen Brain Atlas API.** Программное обеспечение содержит методы доступа к изображениям с высоким разрешением, вычислениям, первичным микрочипам и результатам

секвенирования РНК, а также файлам МРТ и DTI из набора ресурсов атласа Института. А так же **Allen Software Development KIT**- комплект разработки программного обеспечения Allen (SDK), содержащий набор программных библиотек, которые взаимодействуют с API и позволяют пользователям легко читать и анализировать набор данных, представленных в репозиторях Allen [11].

Данный проект делает упор на роль генов в когнитивных расстройствах, стараясь разобраться в причинах патологий посредством атласов. Проект рассматривает деятельность мозга с точки зрения его биологического устройства, но не ставит перед собой целью смоделировать мышление мозга или «машину», способную принимать решения.

Human Connectome Project

Коннектомом называется описание структуры связей в нервной системе. Считается, что в связях между нейронами заключены характеристики человеческой индивидуальности, личность и интеллект,- поэтому описание коннектома человека может стать большим шагом к пониманию многих умственных процессов. Коннектом уникален для каждого человека, также как и геном [13].

Глобальная цель Human Connectome Project (HCP) — получить карту нейронных связей человеческого мозга, для того, чтобы решить многие заболевания психики и нервной системы (аутизм, шизофрения). Это значительно улучшит возможности визуализации и анализа соединений мозга. Возможности проекта предполагают обеспечение беспрецедентной компиляции нейронных данных и интерфейса для графического перемещения этих данных, т.е. участник проекта может сравнивать основные схемы мозга, связи, увеличивать масштаб так, чтобы исследовать клетки. Себастьян Сеунг, руководитель проекта, говорит о том, что сами по себе нейроны здоровы, но они, возможно, неверно закоммутированы, т.е.

патологии вызваны наличием в мозгу неправильных связей. На рисунке 6 представлено исследование мозга с помощью МРТ.

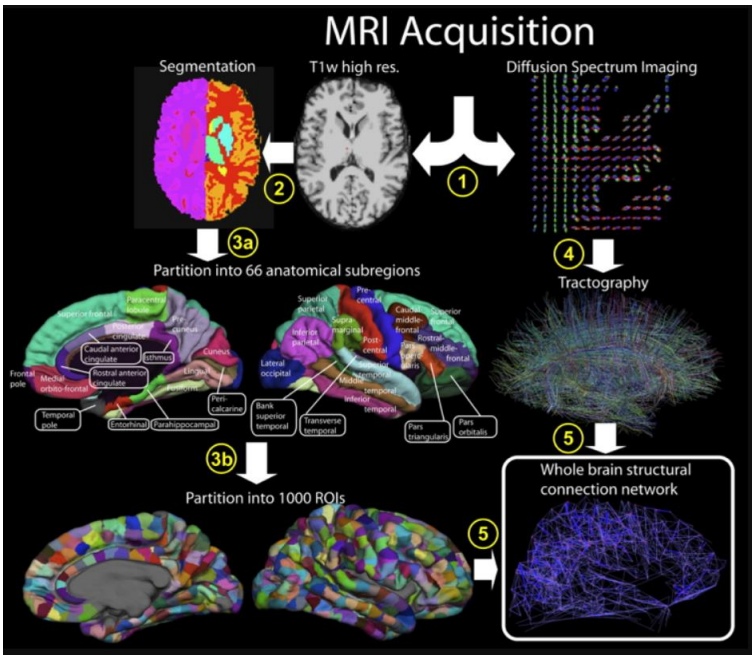

Рис. 6. Исследование связей мозга человека с помощью МРТ

Первым среди живых существ был описан коннектом червя Caenorhabditis elegans в 1986 году [14]. Его нервная система насчитывает всего 302 нейрона и порядка 7000 соединений. Определение коннектома червя-нематоды заняло более 12 лет. Для мозга человека, который содержит примерно 100 млрд. нейронов и около 10тыс. соединений, эта задача кажется невозможной, но HCP старается сделать невозможное возможным, над этим работают специалисты разных областей: биологи, врачи, инженеры, компютерные специалисты, физики и др. Некоторые из них сканируют мозг (с помощью МРТ) 1200 здоровых взрослых, чтобы создать карту мозга с высоким разрешением, другие исследуют генетическую и поведенческую информацию для построения более полной картины нейронной архитектуры мозга. В ходе исследования учёные сопоставили полученные трёхмерные модели мозга с анкетными данными испытуемых: IQ-тест, социоэкономические показатели, случаи применения насилия в прошлом и

др. Исследователи пытались определить, есть ли связь между определённой конфигурацией мозговых соединений и определёнными характеристиками людей [15]. На рисунке 7 представлен коннектом червя нематоды.

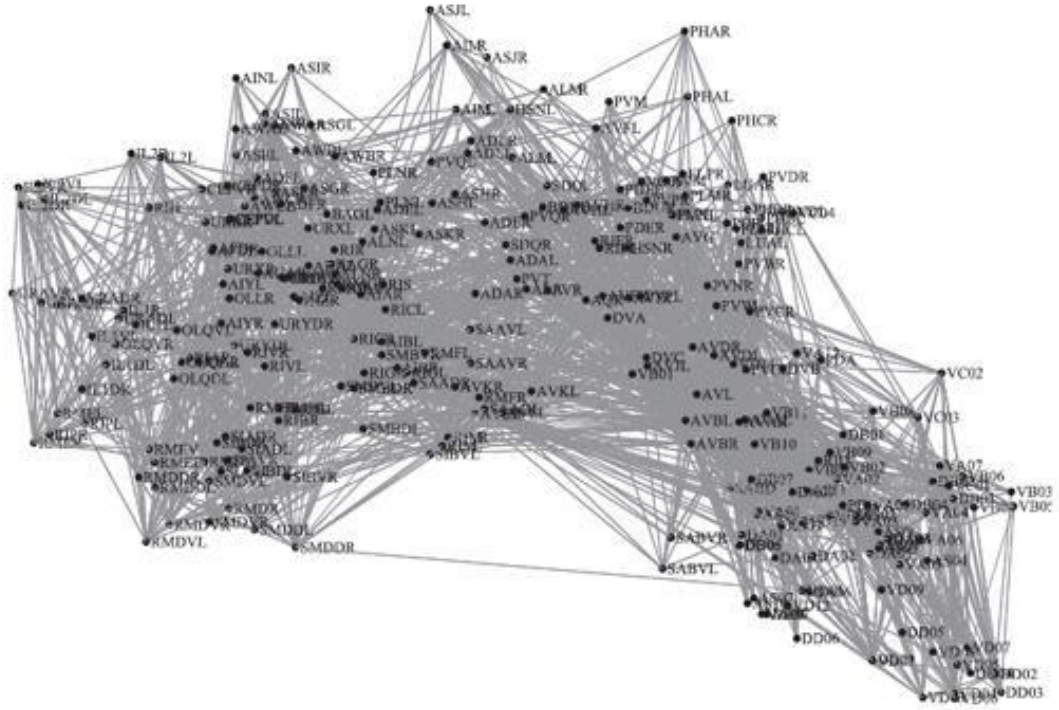

Рис. 7. Коннектом червя-нематоды

Когнитивные системы. IBM Watson

Мозг человека – мощная система, которая способна анализировать неструктурированные массивы данных, обрабатывать и приводить к структурированному виду. Каждый день человечество генерирует около 2,5 квинтиллионов байтов данных, и 80% из них являются неструктурированными, и большинством компьютерных систем эта информация не учитывается [16]. Для того чтобы обрабатывать большие объемы неструктурированных данных с помощью компьютеров, создали когнитивные вычисления, которые частично повторяют особенности работы мозга человека, но способные работать намного быстрее и эффективнее (во

внимание берется малая часть функций мозга, ответственных за анализ и обработку поступающей извне информации, а так же за самообучение).

На рисунке 8 представлены основные характеристики когнитивных систем.

Рис. 8. Основные характеристики когнитивной системы

Мощнейшим суперкомпьютером, основанном на когнитивных вычислениях является анонсированный в 2011 году Watson компании IBM. Watson не искусственный интеллект – это мощная экспертная система, задача которой отвечать на вопросы человека, заданные на естественном языке. Система не использует заранее подготовленные ответы, а находит их самостоятельно, основываясь на приобретенных знаниях, и дает оценку достоверности этих ответов [17, 18].

Для обучения системы сложному анализу, учитывая эмоции и контекст, специалисты используют глубокую обработку естественного языка – вопросно-ответную систему контентной аналитики Deep Question Answering (DeepQA) [19]. При обработке информации устанавливаются связи и корреляции между различными данными, фактами, событиями и явлениями. Одна из главных задач системы — выявление связей, которые незаметны простому глазу и которые не могут быть выявлены обычным

способом, либо затрачивают слишком много времени на поиск [20]. Так, Watson способен составить план лечения мозга больного раком за 10 минут, это в 960 раз быстрее, чем его составляет врач [21].

В 2013 году Watson стал использоваться в коммерческих целях в качестве диагноста. На протяжении двух лет суперкомпьютер изучал 605 тыс. медицинских документов, около 2 миллионов страниц текста, проанализировал 25 тыс. историй болезни и проработал 15 тыс. часов для настройки алгоритмов.

Компьютер IBM Watson назначает оптимальное лечение с точностью 90%, что значительно превосходит человека в точности диагностирования отдельных видов рака и назначения лечения.

Показательный случай произошёл в Японии. После того, как лечение женщины, страдающей лейкемией, оказалось неэффективным, команда японских врачей обратилась за помощью к IBM Watson, которая смогла успешно определить, что она действительно страдала от другой, редкой формы лейкемии, чем первоначально считали врачи. Watson, чтобы поставить диагноз, сравнил генетические данные и историю болезни пациентки с информацией из 20 млн. других историй болезни в своей базе и предложил другое лечение [22].

Рэймонд Курцвейл, американский изобретатель и футуролог, известный созданием систем распознавания речи и исследованием искусственного интеллекта, сказал об этом так: «Много писали о том, что Watson работает через статистические расчеты, а не через "подлинное" понимание. Некоторые понимают это так, что Watson просто собирает статистику о словосочетаниях, но точно так же и пространственное распределение концентрации нейротрансмиттеров в коре человеческого мозга можно назвать "статистической информацией". В самом деле, мы разрешаем свои сомнения примерно так же, как это делает Watson, — сравнивая вероятности различных интерпретаций фразы» [23]. По мнению

же философа Джона Сёрль, Watson не компьютер, который умеет думать, это обученная система поиска ответов.

Помимо использования суперкомпьютера в медицине, разработчики нашли применение и в других сферах: маркетинг, финансы, логистика, аудит и даже кулинария. Примечательно, что на данный момент IBM Watson не способен анализировать русскоязычные данные. Кроме того, менталитет русского человека сильно отличается от других, поэтому прогнозировать корректную работу программы для России сложно. При этом, учитывая большую вычислительную мощность, хорошее распознавание естественного языка человека и точность результатов суперкомпьютера, речь о моделировании области мозга человека здесь не идет, хоть и деятельность когнитивных систем тесно связана с функциями головного мозга (ответственных за анализ, обработку и самообучение).

Разработчики Watson не учитывали при создании когнитивной системы особенности мышления мозга человека в нестандартных ситуациях: экстремальных ситуациях, люди сразу замечают вещи, которые привлекают их внимание, творческий процесс мысли, психофизические особенности, творческое мышление, менталитет, особенности гражданского общества и социума, которые влияют на мыслительный процесс и восприятие действительности, а так же особенности двух полушарий мозга, отвечающих за разные мыслительные процессы, по разному реагирующие на действительность и принятия решений.

Помимо того, что работа мозга сама по себе вещь неизведанная, работа мозга русского человека, со своими конкретными нейронными связями, образующимися под воздействием социума, традиций и т.д. не рассматривается в условиях моделирования деятельности мозга. Мы делаем акцент на обучении системы на реальных данных, полученных от исследований людей, живущих в России, подготовленных Санкт-Петербургским научно-исследовательским психоневрологическим институтом им. В. М. Бехтерева.

Исследования в области мозга приобретают глобальные масштабы, вышеописанные проекты пытаются ответить на вопросы, которые во многом определяют нас: как работает наш мозг, сознание, в чем причины психических расстройств и нейродегенеративных заболеваний, большинство проектов исследуют мозг с биологической точки зрения, а так же с точки зрения взаимодействия нейронов.

Сейчас подобные исследования – крупные наднациональные проекты, но для их успешной реализации необходима эффективная инфраструктура (банки образцов, масштабные производства средств для научных исследований и т. п.), огромная вычислительная мощность, миллионное и даже миллиардное финансирование.

Глава 1. Особенности асимметрии головного мозга человека

Межполушарная асимметрия мозга играет важную роль в жизни индивида и всего человечества и проявляется не только на уровне анатомических особенностей, но и в асимметрии психических процессов. При доминировании одного полушария (левого или правого) наблюдается функциональная межполушарная асимметрия головного мозга, что способствует определению ведущей руки (правша/левша), способов мышления, эмоционального восприятия, воображения и пр. [24]. При этом оба полушария имеют способность работать относительно независимо друг от друга, перераспределяя функции между собой – *латерализация* функций мозга [25, 26].

Основы функциональной специализации полушарий являются врожденными, однако по мере развития ребёнка и влияния множества биосоциальных факторов происходит усложнение и совершенствование механизмов межполушарного взаимодействия, что приводит к *латеральности* – формированию определенного профиля межполушарной асимметрии. Степень выраженности асимметрии является продуктом онтогенетического развития [27, 28].

1.1. Экспериментальное подтверждение асимметрии мозга

Роджер Уолкотт Сперри — американский нейропсихолог, профессор психобиологии, получивший в 1981 году Нобелевскую премию по физиологии и медицине «За открытия, касающиеся функциональной специализации полушарий головного мозга», объяснил, что познавательные функции двух полушарий мозга во много различаются [29]. Р. Сперри провёл

эксперимент над пациентами, страдающими эпилепсией: эпилептические припадки ослабевают или вовсе прекращаются, если рассечь мозолистое тело, которое связывает между собой оба полушария и является их коммутатором (рис. 9).

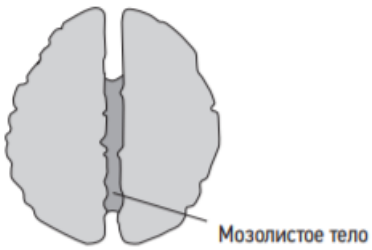

Рис. 9. Мозолистое тело головного мозга

После проведения данной операции у здоровых пациентов наблюдались несвойственные им ранее симптомы, например, некоторым стало крайне трудно сложить кубики определенным образом. В другом эксперименте Сперри пациенты с разделенными полушариями головного мозга садились перед экраном так, чтобы экран заслонял их руки. Пациент смотрел в центр экрана, на котором слева появлялось слово, данная информация передавалась в правое полушарие (рис. 10). Выяснялось, что человек не может произнести слово, которое видел на экране несколько секунд назад. Затем пациент левой рукой, скрытой за экраном, должен был выбрать предмет, название которого было написано на экране, при этом пациент выбирал правильный предмет, не осознавая того, что видел какое-то слово (правое полушарие отвечает за движения левой стороны тела). Этот опыт показывает, что речь и умение читать формируется в левом полушарии, правая часть обрабатывает визуальные образы, т.е. разная информация обрабатывается разными полушариями [30].

Рис. 10. Эксперимент Р. Сперри. [30]

Еще одно подтверждение асимметрии показала ангиография сонных артерий (рентген мозга после введения в сонную артерию контрастного вещества), произведенная Нормандом Гешвиндом – американским неврологом. Эксперимент показал, что сильвиева борозда (латеральная), которая отделяет височную долю в коре головного мозга, в левом полушарии длиннее и более прямая, чем в правом [31, 32]. На рисунке 11 наглядно представлены анатомические различия левого и правого полушарий мозга.

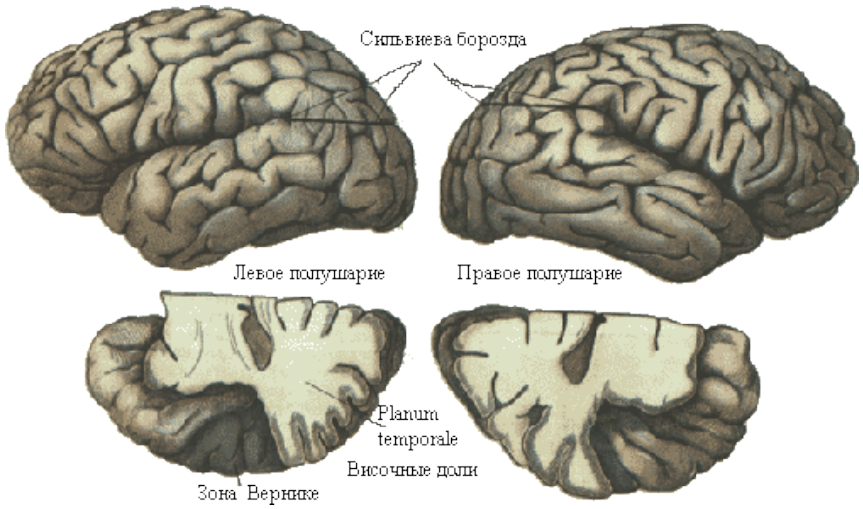

Рис. 11. Анатомическая асимметрия полушарий мозга [33]

Левое полушарие информацию обрабатывает последовательно и аналитически. Основными областями, за которые оно отвечает, являются: вербальные операции, обработка временных взаимоотношений, математические расчеты, абстрактное мышление, интерпретация

символических понятий, формирование речевых функций, а также управление функционированием правой стороны тела [30, 34]. Левое полушарие преимущественно использует индуктивное мышление, т.е. вначале анализирует информацию, а затем осуществляет синтез, получая частное заключение из общего [35]. **Правое полушарие** обрабатывает информацию одновременно и интуитивно, осуществляя индуктивный способ мышления (сначала осуществляются процессы синтеза, затем анализа информации). Оно лучше, чем левое, справляется с задачами распознавания зрительных образов и пространственных взаимоотношений, эффективно интерпретирует сложные взаимосвязи, звуковые образы (голос, интонацию) и «понимает» музыку, а так же управляет функционированием левой стороны тела [29]. Понимание метафор, эмоций, форм, творческий подход к решению задач относят также к управлению правого полушария [30]. Кроме того, по мнению Р. Сперри оба полушария обладают способностью к сознанию и самосознанию, а также осознанию взаимоотношений в социуме [36].

1.2. Физиологические особенности головного мозга

Стоит понимать, что головной мозг человека не разделить только на левое и правое полушарие, не ставя во внимание всю его морфологию. Рассмотрим строение головного мозга человека немного подробнее.

Головной мозг – это отдел нервной системы, находящийся в черепной коробке, который состоит из мозгового ствола, мозжечка и конечного мозга (рис. 12). Внутри головного мозга имеются желудочки, заполненные спинномозговой жидкостью.

Ствол мозга включает продолговатый мозг, средний мозг, промежуточный мозг и варолиев мост. Центры дыхания, сердечной деятельности, обмена веществ, чувствительные и двигательные центры

расположены в стволе мозга. Здесь же находится нервное образование – ретикулярная формация, которая регулирует тонус мозговой коры.

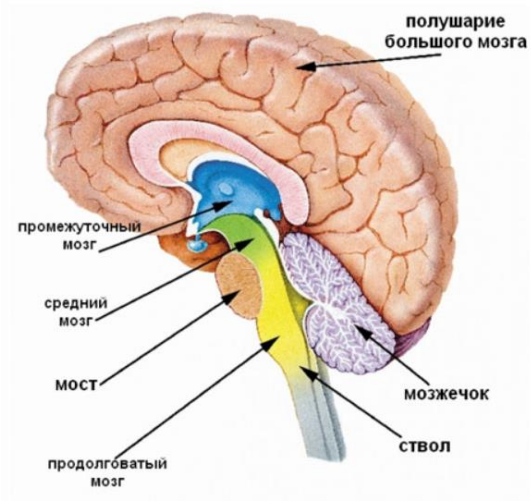

Рис. 12. Отделы головного мозга [37]

Мозжечок ответственен за координацию движений, участвует в поддержании тонуса мышц, в проявлениях вегетативной нервной системы.

Конечный мозг состоит из базальных узлов и больших полушарий. Базальные узлы имеют центры регуляции двигательных автоматизмов, сенсорных функций, вегетативных проявлений, участвуют в процессах речи, психики человека. Из всех отделов головного мозга наиболее развиты большие полушария [38].

Большие полушария покрыты корой, – слоем серого вещества толщиной около 3 мм и соединены пучками нервных волокон. В передней части полушарий выделяют лобную долю, две теменные в верхней, в боковой – две височные, и затылочную – в задней части полушарий (рис. 13). Лобные доли отвечают за способность выносить суждения, височные доли за память, слух и стабильность настроения, теменные за обработку информации от органов чувств осязания обоняния и вестибулярного аппарата, затылочные доли обрабатывают визуальную информацию.

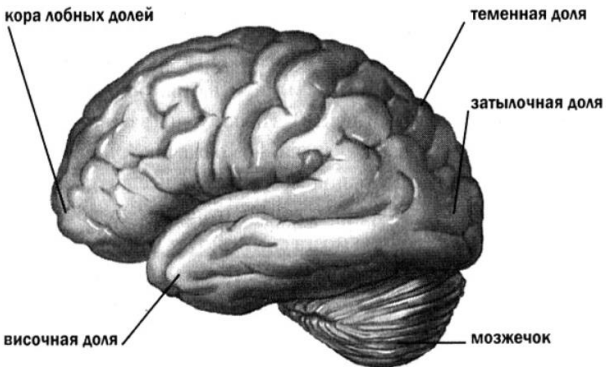

Рис. 13. Доли больших полушарий мозга

В головном мозге также находятся: передняя часть поясной извилины, базальные ганглии, которые отвечают за контроль удовольствия и беспокойства и лимбическая система, служащая эмоциональным центром. На рисунке 14 показано правое полушарие и его «составляющие».

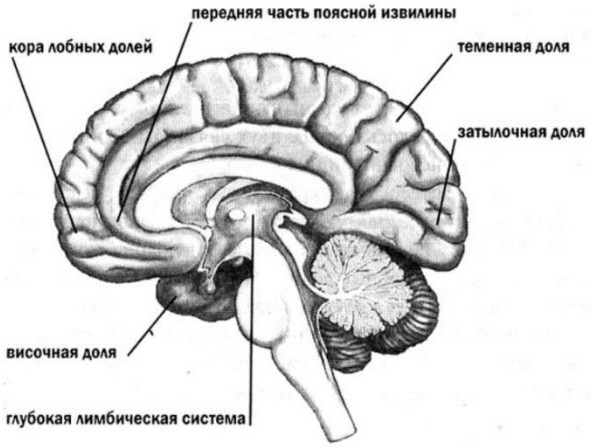

Рис. 14. Вид правого полушария со стороны срединного разреза

Иными словами, задняя часть мозга (затылочные доли, теменные и задняя часть височных долей) ответственны за восприятие мира. В передней части происходит синтез и анализ информации, принятие решений, планирование и процесс, ответственный за реализацию намерений [39].

В функциональном отношении левое и правое полушария неравнозначны. При отключении правого полушария интеллектуальный уровень человека остается неизменным, но повышается позитивный

эмоциональный фон. При отключении левого проявляется депрессия. Оба полушария как бы сдерживают друг друга, нормализуя общую деятельность головного мозга [38].

1.3. Мозг и психика человека

Важно принимать во внимание особенность деятельности мозга каждого индивида в зависимости от психики человека. Различные структуры нервной системы участвуют в обеспечении психических процессов**.** Советский психолог, основатель отечественной нейропсихологии Александр Романович Лурия выделил три основных функциональных блока мозга, которые присутствуют при любой психической деятельности [41].

Блок регуляции тонуса и бодрствования, за который отвечает *ретикулярная формация*, поддерживающая активное состояние нервного аппарата. Данный блок обеспечивает решение различных задач, пробуждает организм, обостряет чувствительность.

Блок приема, переработки и хранения информации. Этот блок располагается в задних отделах коры головного мозга и включает височную, теменную и затылочную доли. Данный блок состоит из надстроенных друг над другом корковых зон трех типов: проекционные зоны, проекционно-ассоциативные зоны, зоны перекрытия. Импульсы с периферии поступают в *проекционные зоны*, в следующей зоне получаемая информация перерабатывается, а *в зонах перекрытия* протекают сложные формы психической деятельности.

Блок программирования, регуляции и контроля деятельности. Структуры блока расположены в передних отделах полушарий. Особенностью данного отдела головного мозга является мощная система связей с другими отделами. Лобные доли получают импульсы от первого функционального блока, получая энергию и тонус. Так же лобные доли регулируют ретикулярную формацию. Здесь находятся и центры речи, а

высшие психические процессы формируются на основе речи. Повреждение лобных долей приводит к нарушению программ поведения человека. Блок отвечает за формирование целей, планов, регулирование деятельности, контроль и осознание ошибок человеком.

Совместная работа вышеописанных функциональных блоков мозга является необходимым условием осуществления любой функции психики человека [42].

Далее представлены результаты исследования, проведенного командой из Новосибирского государственного медицинского университета и Новосибирского гуманитарного университета, в состав которой вошли доктор медицинских наук Куликов В.Ю., кандидат медицинских наук Антропова Л.К., врач-невролог Козлова Л.А. и кандидат психологических наук Андронникова О.О. Исследование способствует пониманию особенностей протекания сложных психических процессов и межполушарной специализации мозга у студентов г. Новосибирска[1].

Исследования взаимосвязи асимметрии полушарий головного мозга с эмоциональными состояниями человека привели к выводу, что оба полушария задействованы в эмоциональной составляющей жизнедеятельности человека. Однако правое полушарие в большей степени связано с негативными эмоциями. Обладатели доминирующего правого полушария чаще интроверты и имеют признаки психопатологического депрессивного синдрома, предрасположенность к стрессу, пассивно оборонительному типу общения, тревожности, зажатости, раздражительности, неустойчивости эмоционального состояния и настроения, вследствие чего способны проявлять неожиданные для себя и окружающих поступки. Обладатели доминирующего левого полушария

[1] Результаты опубликованы в журнале «Медицина и образование в Сибири», выпуск № 3, 2011 г. авторы Л.К. Антропова, О.О. Андронникова, В.Ю. Куликов.

экстраверты, редко склонны к депрессии, эмоционально устойчивы, обладают высоким уровнем устойчивости, самоуверенностью и силой воли.

В условиях стресса человек с леволатеральным профилем старается достичь позитивного изменения ситуации, в таком случае человек менее зависим от социума и чужого мнения. Человек с доминирующем правым полушарием, напротив, в стрессовой ситуации избегает проблему психически и физически [43].

Таким образом, можно сделать вывод, что левостороннее доминирование функциональной межполушарной асимметрии является линейным, логическим, аналитическим и лингвистическим. Правостороннее невербальное, образное, чувственное. Левое полушарие преимущественно «цифровое», которое разбивает информацию по категориям «включено/выключено», «1/0» , «true/false». Правое полушарие – скорее «аналоговое» [44].

1.4. Влияние эмоций на принятие решений

С. Брёйгельманс и М. Зееленберг – профессора отделения Социальной психологии Университета Тилбурга в Нидерландах, исследовали влияние эмоций и чувств на принятие решений и поведение человека. Результаты привели к выводу, что эмоции мотивируют определенное поведение; разные эмоции обладают разным воздействием; понимание эмпирического содержания эмоций позволит предсказывать будущее поведение [45, 46].

Человеческий мозг часто сравнивают с мощной вычислительной машиной, но человек не производит в уме сложные математические вычисления и не рассчитывает вероятность, когда принимает какие-либо решения. Мы приспособлены принимать решения быстро, часто не задумываясь о том, почему сделали такой выбор. Человек мыслит эвристически. Важным видом эвристики является аффект – внезапный эмоциональный порыв, в данном случае человек однозначно понимает, какие

чувства испытывает и может сделать выбор в пользу того, что нравится. Например, слово «теракт» вызывает негативные эмоции, а словосочетание «лазурный берег» воспринимается положительно. Этот автоматический одномерный импульс скрывает любой риск или пользу как самостоятельные величины, если человеку что-то очень нравится, он старается минимизировать своё восприятие риска.

Эксперимент, который проводился в период с 1982 по 1997 год Дэвидом Хиршлейфером и Тайлером Шамвей подтверждает зависимость принятия решений от эмоциональной составляющей. Исследователи изучали, как утренний солнечный свет влияет на биржевую активность. В наблюдение попали 26 бирж, и тенденция прослеживалась однозначно: если с утра светит солнце, то акции на бирже растут, т.к. с утра люди способны к принятию важных решений и обдумыванию сложных задач, они полны сил и готовы к новому дню, а солнечный свет вызывает положительные эмоции, а следовательно влияет на совершаемые действия [46].

1.5. Особенности менталитета

Исследовано участие полушарий мозга в зависимости от принадлежности индивидуума к различным культурам, кроме того, выявлено, что в процессах социальной адаптации большую роль принимает левое полушарие, а в физической адаптации правое [47, 48].

Кроме того, различные народы обладают отличительными чертами психотипов, так к преобладающему психотипу западного народа относят следующие черты:

- техницизм,
- тип мышления: левополушарный,
- способность к аналитике больших объёмов информации,
- преобладание разума над чувствами,
- склонность к оптимизму, лидерство, активность,

- рациональность,
- материалистическая ориентация.

К психотипу восточного человека относят:

- мистицизм, вера,
- иррациональность,
- тип мышления: правополушарный,
- преобладание чувств над разумом,
- склонность к пессимизму, пассивности, консерватизму.

Специалисты психологии и соционики в русском психотипе выделяют:

- большая доля иррациональности,
- интуиция, чувствительность, созерцание,
- преобладание чувств над разумом,
- эмоциональность,
- преобладание индивидуально-личностных отношений над формальными,
- мораль выше законов,
- антиматериалистическая ориентация [49, 50, 51].

«Умом Россию не понять», как говорил Фёдор Иванович Тютчев.

Глава 2. Выбор компонентов установки

2.1. TrueNorth

Микрочип, имеющий не-фон Неймановскую нейроморфную архитектуру, TrueNorth разработан компанией IBM в рамках программы DARPA SyNAPSE, обеспечивает глубокое обучение с высокой энергоэффективностью, скоростью и масштабируемостью (рис. 15). Проект имеет симулятор чипа, язык программирования Corelet и программное обеспечение Compass для разработки программ к чипу TrueNorth [52].

Рис.15. Чип TrueNorth от IBM

Архитектура фон Неймана предусматривает наличие процессора и памяти. Обмен данными между процессором и памятью происходит с ограничением производительности по специальному каналу, что является «бутылочным горлышком». Выполнение операций последовательное, зависимо от тактового генератора, чем больше частота тактов, тем выше скорость обработки данных, а, следовательно, и энергопотребление процессора [53].

Архитектура TrueNorth схожа с неокортекстом. Биологический мозг не разделяется на процессор и память, есть сигналы между нейронами, которые передаются друг другу с помощью синапсов. Подобная структура является динамической, обработка данных происходит параллельно, обработка сигналов, коммуникации и память находятся на одном чипе, не требует больших тактовых частот, энергоэффективность крайне высока. Система позволяет объединять чипы, наращивая вычислительную мощность нейроморфных систем. TrueNorth соответствует деятельности правого полушария, способен обрабатывать большое количество сенсорных данных, распознавать образы в реальном времени и др. [54].

Технические характеристики:

- 5,4 млрд. транзисторов,
- 4096 параллельных и распределенных ядер, взаимосвязанных в сети на чипе в массиве 64x64, 1 млн. нейронов, 256 млн. синапсов (рис. 16),
- масштабируется до 65 536 ядер, 16 чипов и 16 миллионов нейронов, 4 млрд. синапсов,
- энергопотребление: 70-100 милливатт [55].

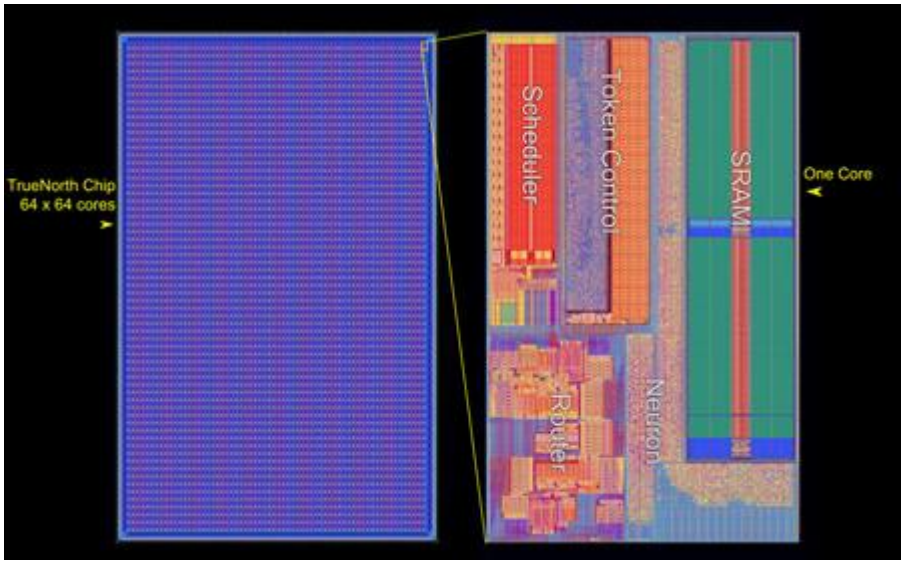

Рис. 16. Основной массив 64x64 чипа TrueNorth

Коммуникации TrueNorth основаны на каналах «точка-точка», передающих сигналы от одного нейрона к одному нейросинаптическому ядру, где эти сигналы могут соединяться с любым другим нейроном.

На рисунке 17 самый левый нейрон в ядре 1 соединяется с ядром 3. Чтобы подключиться к ядрам 2 и 3, 2-й и 3-й нейроны в ядре 1 дублируют друг друга и каждый из них делает одно соединение [56].

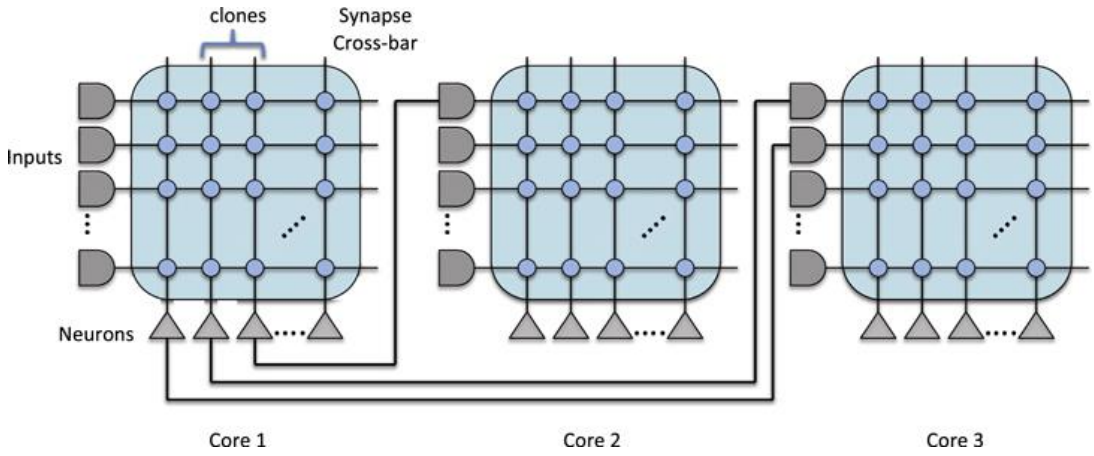

Рис. 17. Связи нейронов в TrueNorth

2.1.1 Системы на базе TrueNorth

Первая платформа с чипом TrueNorth – NS1e от IBM, способная эмулировать 1 млн. нейронов (рис. 18).

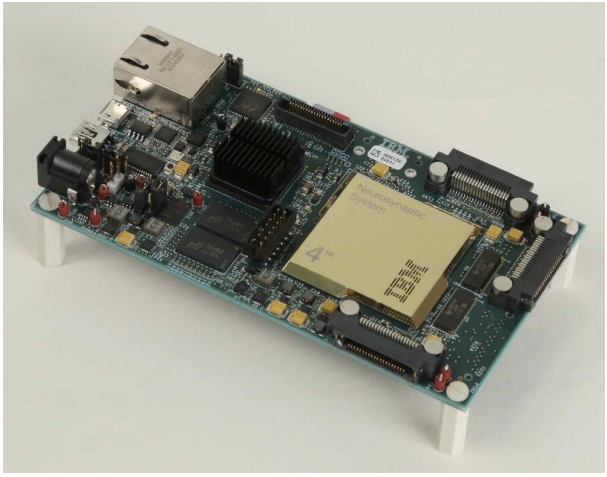

Рис. 18. Плата NS1e.

Следующей стала NS1e-16, представляющая собой масштабируемую систему из 16 плат NS1e с блоком питания, коммутатором Ethernet и сервером Linux, все из которых размещены в компактном автономном блоке, где каждая из плат NS1e может быть легко интегрирована, но смонтирована таким образом, чтобы платы могли легко меняться местами (рис. 19) [57].

Рис. 19. Плата NS1e-16.

Третий этап – NS16e. NS16e имеет 16 чипов TrueNorth, которые объединены в одной плате, что эквивалентно примерно 16 миллионам нейронов и 4 миллиардам синапсов (рис. 20).

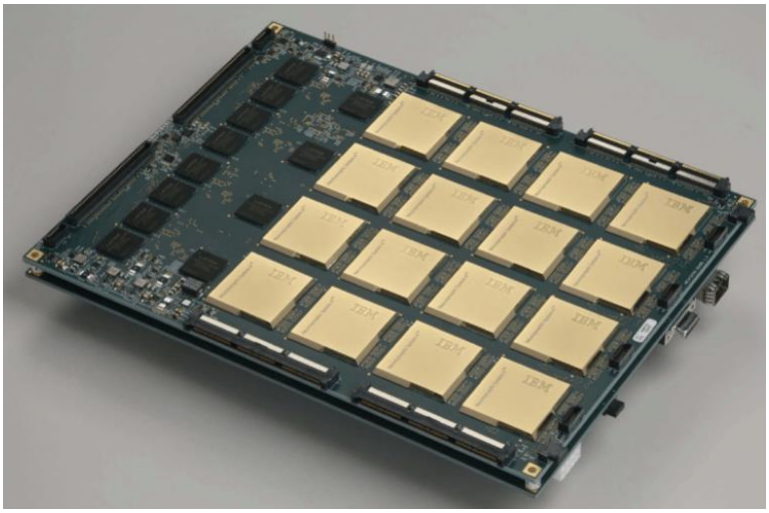

Рис. 20. Плата NS16e, включающая 16 чипов TrueNorth

Устройство может сохранять информацию и принимать решения на основе шаблонов, обнаруженных через вероятности и ассоциации. Используя обучающие модели и алгоритмы, компьютер может связывать прошлые и текущие данные с шаблонами и классификациями. В некотором смысле система работает по принципам алгоритмов машинного обучения, распознавания образов. Однако система IBM отличается тем, что использует NS16e, чтобы приблизить понимание, как работают нейроны и синапсы мозга [58].

2.2 POWER9

Процессор POWER9 имеет RISC-архитектуру[2]. Компания IBM выделяет несколько областей применения POWER9:

- прогнозирование, искусственный интеллект (ИИ), когнитивные вычисления;
- обработка больших данных, в т.ч. в облаке;
- задачи корпоративных вычислений;
- High Performance Computing (HPC) – высокопроизводительные вычисления [59].

POWER9 – 14-нанометровый процессор, содержащий 8 млрд. транзисторов (площадь 695 кв. мм), частоту почти 4 Ггц, имеет более высокую производительность одного потока, в сравнении с предыдущим поколением Power8 [59].

Структура процессора разделена на внешний компонент и компонент блоков выполнения (EU). На уровне EU Slice (S) находятся векторно-скалярные устройства (VSU), которые работают с 64-разрядными данными,

[2] Reduced instruction set computer – «компьютер с сокращённым набором команд». Увеличение быстродействия за счёт упрощения инструкций, упрощая их декодирование и уменьшая время выполнения.

выполняются целочисленные операции, операции с плавающей запятой, а также криптографические команды. В структуре имеются блоки загрузки/записи данных (LSU), устройство выполнения переходов (BRU) и устройство выборки команд (IFU) (до восьми команд за такт в буфере и оптимизированный предсказатель переходов для поддержки внеочередного выполнения команд) [59, 60].

Суперслой (SS), образованный из двух Slice, работает со 128-разрядными данными. SMT4 – процессорное ядро (аппаратная поддержка четырех программных нитей) основывается на нескольких суперслоях с содержанием традиционных блоков (рис. 21). POWER9 поддерживает и ядра SMT8, включающие четыре 128-разрядных SS и увеличенные традиционные блоки. Максимальное количество процессорных ядер SMT4 в микросхеме 24, в SMT8 — 12. SMT4 способна завершать до 128 команд на каждом такте, а SMT8 до 256 [60].

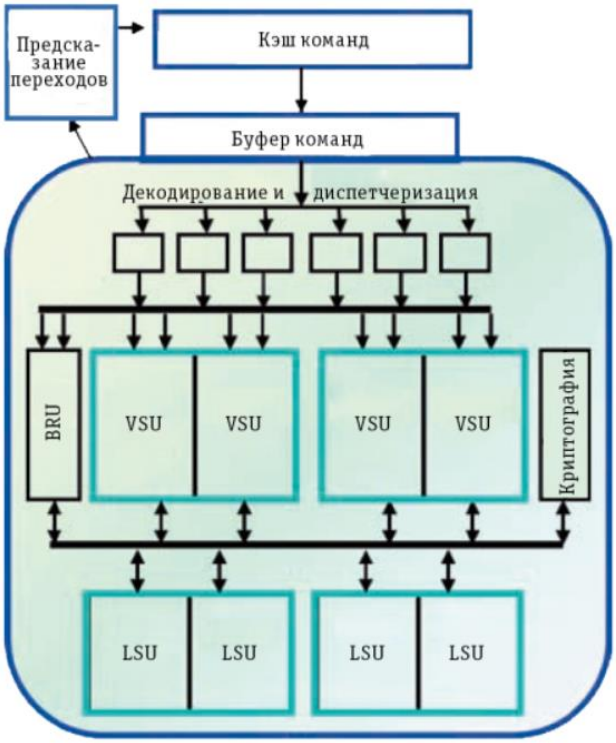

Рис. 21. Архитектура ядра SMT4

Организация кэша:

- L1I Cache
  - 32 Кбайт, 8-канальный наборно-ассоциативный
  - SMT4
- L1D Cache
  - 32 Кбайт, 8-канальный наборно-ассоциативный
  - SMT4
- L2 Cache
  - 258 Кбайт
  - SMT4
- L3 Cache
  - 120 Мбайт eDRAM
  - Двумерная топология 9*12 элеметнов 10 MiB 20-канальный наборно-ассоциативный
  - 7 TB/s общая пропускная способность [60].

Оперативная память POWER9 основана на технологии DDR4 с содержанием двух групп вычислительных SMP-систем: «горизонтальное» и «вертикальное» масштабирование. Речь идет об одно-, двух- или многопроцессорных SMP-системах. Для горизонтального масштабирования память построена с обычным прямым подсоединением через восемь портов DDR4 с общей пропускной способностью 120 Гбайт/с. Максимальная емкость памяти на один процессорный разъем 4 Тбайт. Для многопроцессорных SMP-систем второго типа работа с памятью организована использованием буферной архитектуры памяти. На рисунке 22 представлен вид общий архитектуры POWER9 [59, 60].

POWER9 предлагает два набора приложений для ускорения: PCIe Gen4, который предлагает 48 полос пропускания с пропускной способностью 192 Гб/с и новую линию 25G, которая предлагает дополнительные 48 полос, обеспечивающие пропускную способность до 300 Гб/с. В дополнение к двум физическим интерфейсам есть набор открытых стандартных протоколов,

которые интегрированы в эти сигнальные интерфейсы. Четыре выдающихся стандарта:

- CAPI 2.0 – POWER9 представляет CAPI 2.0 через PCIe, что в четыре раза превышает пропускную способность, предлагаемую исходным протоколом CAPI, предлагаемым в POWER8.
- New CAPI – новый интерфейс, который работает поверх интерфейса POWER9 25G (300 Гб/с), предназначенного для приложений CPU-Accelerators
- NVLink 2.0 – высокая пропускная способность и интеграция между графическим процессором и процессором (предназначена для GPU Volta).
- On-Chip Acceleration – массив ускорителей, предлагаемых самой архитектурой POWER9 [59, 60].

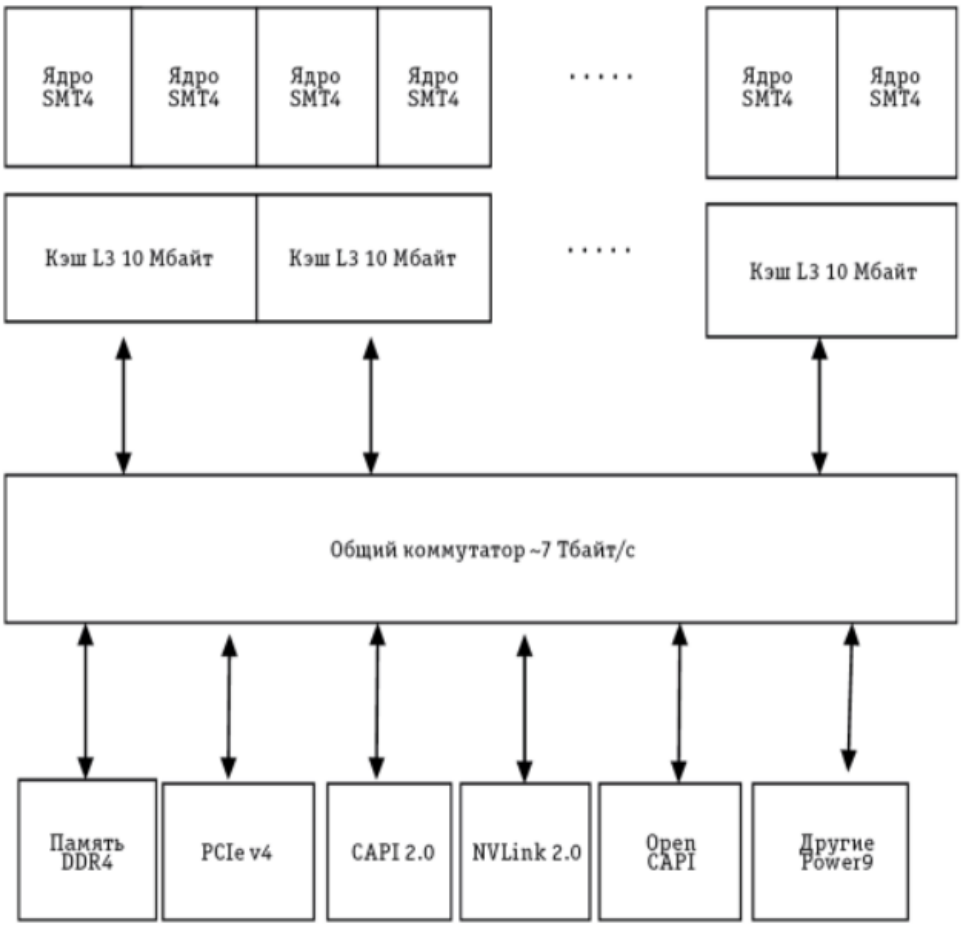

Рис. 22. Общая архитектура POWER9

2.3 GPU NVIDIA Tesla P100 и хост-сервер DGX-1

Зачастую в дата-центрах для поддержания высокопроизводительных вычислений (HPC) создают масштабные кластеры, представляющие собой взаимосвязанные вычислительные узлы CPU, которые могут занимать целые помещения. Графический процессор NVIDIA Tesla P100 с архитектурой Pascal, основанной на базе 16-нм процесса FinFET, имеющего 15,3 млрд. транзисторов, создан для обеспечения высокой энергоффективности и высокой производительности в задачах супервычислений области искусственного интеллекта. К примеру, один серверный узел с двумя Tesla P100, связь которых обеспечивается шиной PCIe, заменяет по производительности около 20 серверных узлов на базе CPU.

Представленный на рис. 23 графический ускоритель хорошо справляется с задачами левого полушария, т.к. его деятельность преимущественно «цифровая», аналитическая.

СПЕЦИФИКАЦИИ ПРОИЗВОДИТЕЛЬНОСТИ УСКОРИТЕЛЕЙ NVIDIA TESLA P100

	Tesla P100 для PCIe серверов	Tesla P100 для серверов с NVLink
Производительность операций двойной точности с плавающей точкой	4,7 Терафлопс	5,3 Терафлопс
Производительность операций одинарной точности с плавающей точкой	9,3 Терафлопс	10,6 Терафлопс
Производительность операций половинной точности с плавающей точкой	18,7 Терафлопс	21,2 Терафлопс
Пропускная способность шины NVIDIA NVLink™	-	160 ГБ/с
Пропускная способность шины PCIe x16	32 ГБ/сс	32 ГБ/с
Полоса пропускания стековой памяти CoWoS с HBM2	16 ГБ или 12 ГБ	16 ГБ
Полоса пропускания стековой памяти CoWoS с HBM2	732 ГБ/с или 549 ГБ/с	732 ГБ/с
Улучшенная программируемость с технологией Page Migration Engine	✓	✓
Защита ECC для повышенной надежности	✓	✓
Оптимизация под сервер для развертывания в дата-центре	✓	✓

Рис. 23. Характеристики Tesla P100

Межсетевое соединение часто регулируется производительностью. Высокоскоростное двунаправленное межсетевое соединение NVIDIA

NVLink предназначено для масштабирования приложений на нескольких графических процессорах, обеспечивая пропускную способность в 5 раз по сравнению с шиной PCIe. На рисунке 24 показано, что GPU Tesla P100 с интерфейсом NVLINK способен обеспечить ускорение приложений до 50 раз [61, 62].

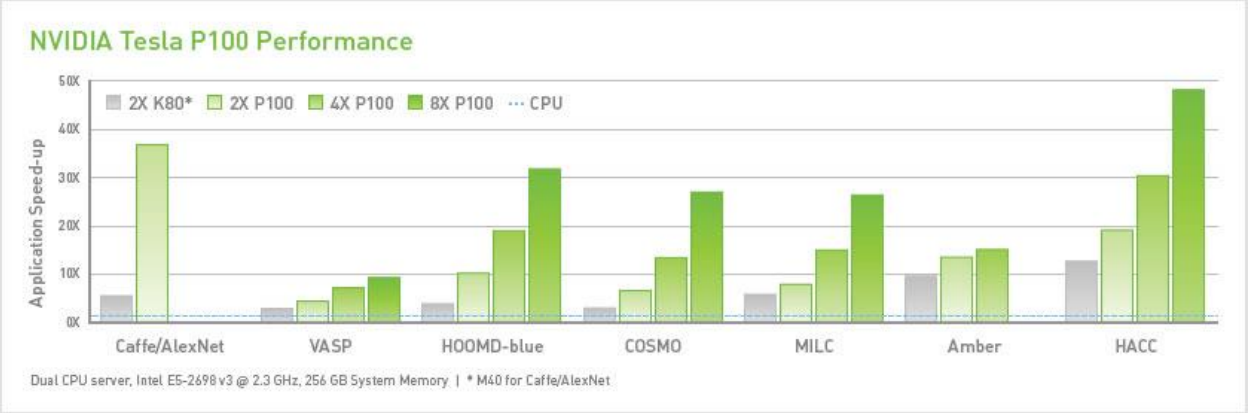

Рис. 24. Ускорение приложений

Для эффективной работы требуется хост-сервер с GPU, таким решением может стать NVIDIA DGX-1 – интегрированная система глубокого обучения. DGX-1 содержит восемь ускорителей Tesla P100 GPU, подключенных через NVLink в гибридной сети кубических ячеек. DGX-1 совместно с двухъядерными процессорами Intel Xeon и четырьмя 100-гигабитными сетевыми интерфейсами InfiniBand обеспечивает высокую производительность для глубокого обучения [61]. Кроме того, системное программное обеспечение DGX-1 и мощные библиотеки настроены для масштабирования глубокого обучения на своей сети графических процессоров Tesla P100. Помимо восьми графических процессоров, DGX-1 включает в себя два процессора для управления загрузкой, управлением хранением данных и глубоким обучением. DGX-1 встроен в корпус с тремя стойками (3U), который обеспечивает питание, охлаждение, сетевое соединение, межсистемное межсоединение и кэш файловой системы SSD,

сбалансированный для оптимизации пропускной способности и глубокого обучения. На рисунке 25 показаны компоненты системы DGX-1 [62].

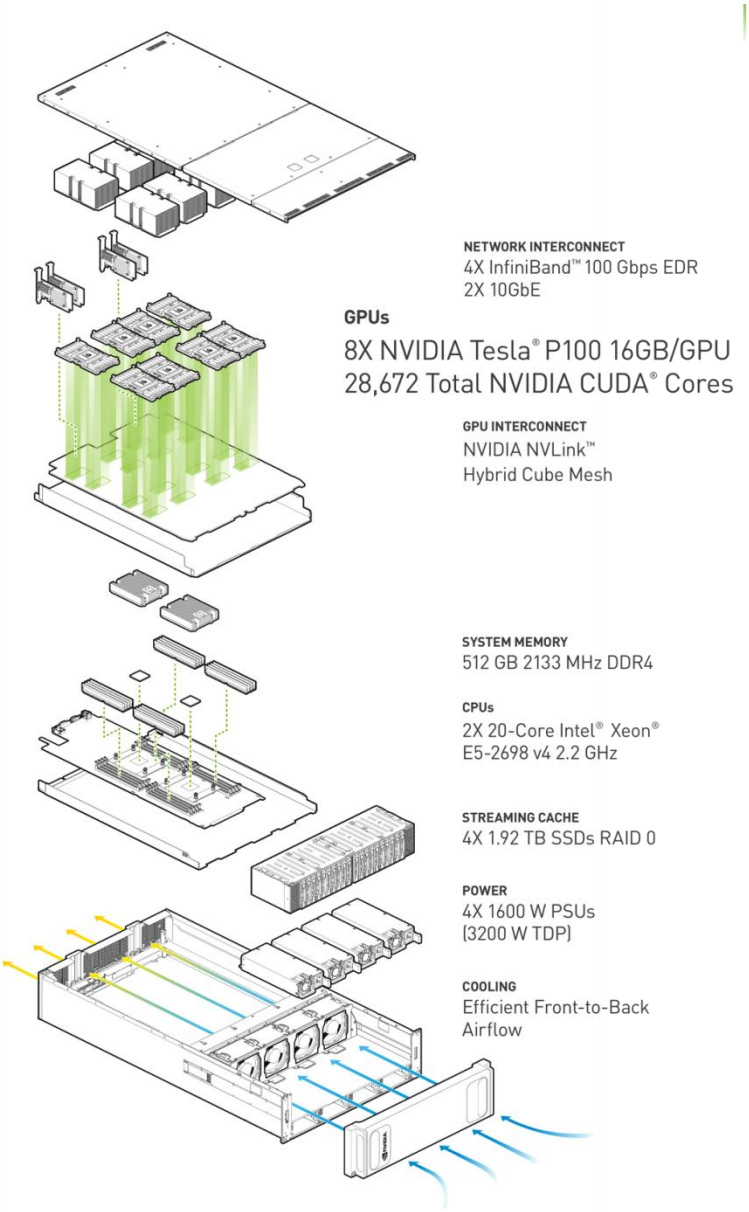

Рис. 25. Комплектация DGX-1

Каждый графический процессор Tesla P100 имеет четыре точки подключения NVLink, каждый из которых обеспечивает двухточечное соединение с другим графическим процессором с максимальной пропускной способностью 20 Gb/sec. Несколько соединений NVLink могут быть соединены вместе, умножая доступную ширину межсоединения между данной парой графических процессоров. NVLink может использоваться для

построения множества сетевых топологий среди множества графических процессоров. Pascal также поддерживает 16 дорожек PCIe 3.0. В DGX-1 они используются для соединения между процессорами и графическими процессорами. PCIe также используется для высокоскоростных сетевых интерфейсных карт (рис. 26) [62].

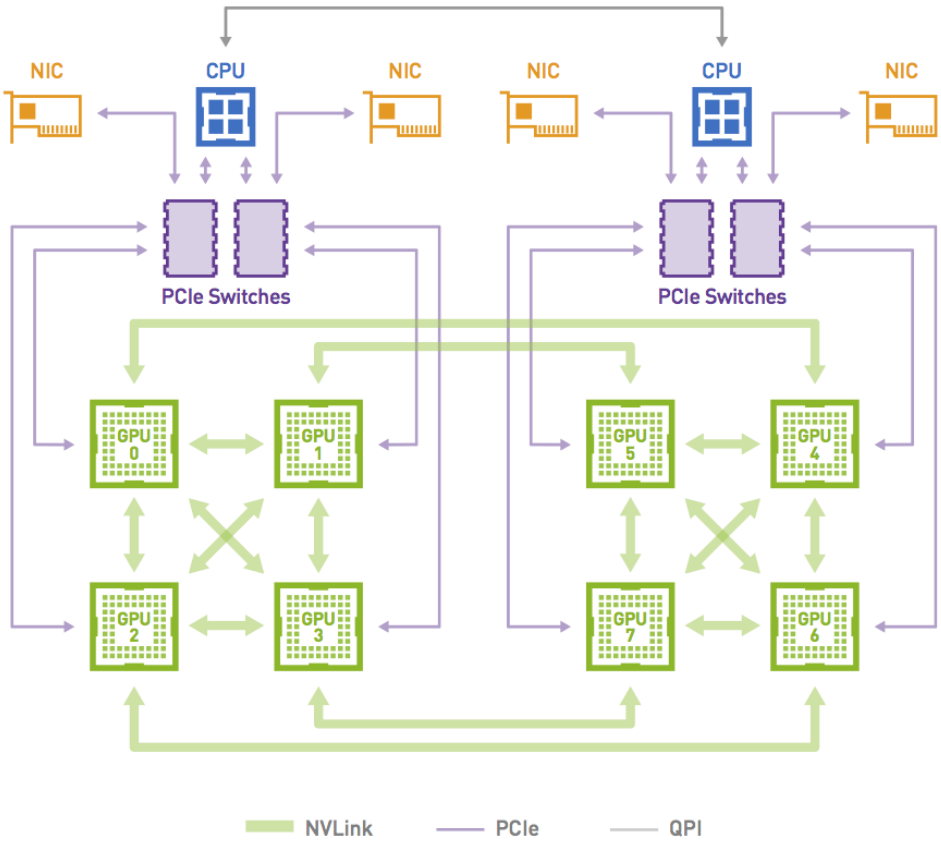

Рис. 26. DGX-1 использует 8-GPU гибридную топологию межсоединений с кубической сеткой. Углы связанных между собой граней куба подключены к сети дерева PCIe, которая также подключается к CPU и сетевым адаптерам.

На рисунке 27 показано ускорение глубокого обучения DGX-1 с использованием всех 8 Tesla P100 из DGX-1 и 8-GPU Tesla M40 и Tesla P100 с использованием межсетевого соединения PCI-e для глубокой нейронной сетевой архитектуры ResNet-50 и Resnet-152 на популярные версии CNTK (2.0 Beta5), TensorFlow (0.12-dev) и Torch (11-08-16). Другое программное обеспечение, задействованное в тестировании: контейнеры NVIDIA DGX

версии 16.12, NCCL 1.6.1, CUDA 8.0.54, cuDNN 6.0.5, Ubuntu 14.04. Драйвер дисплея NVIDIA Linux 375.30. 8x M40 и 8x P100 PCIe – это SMC 4028GR с двумя процессорами Intel Xeon E5-2698v4 и 256 ГБ оперативной памяти DDR4-2133 (DGX-1 имеет 512 ГБ DDR4-2133) [62].

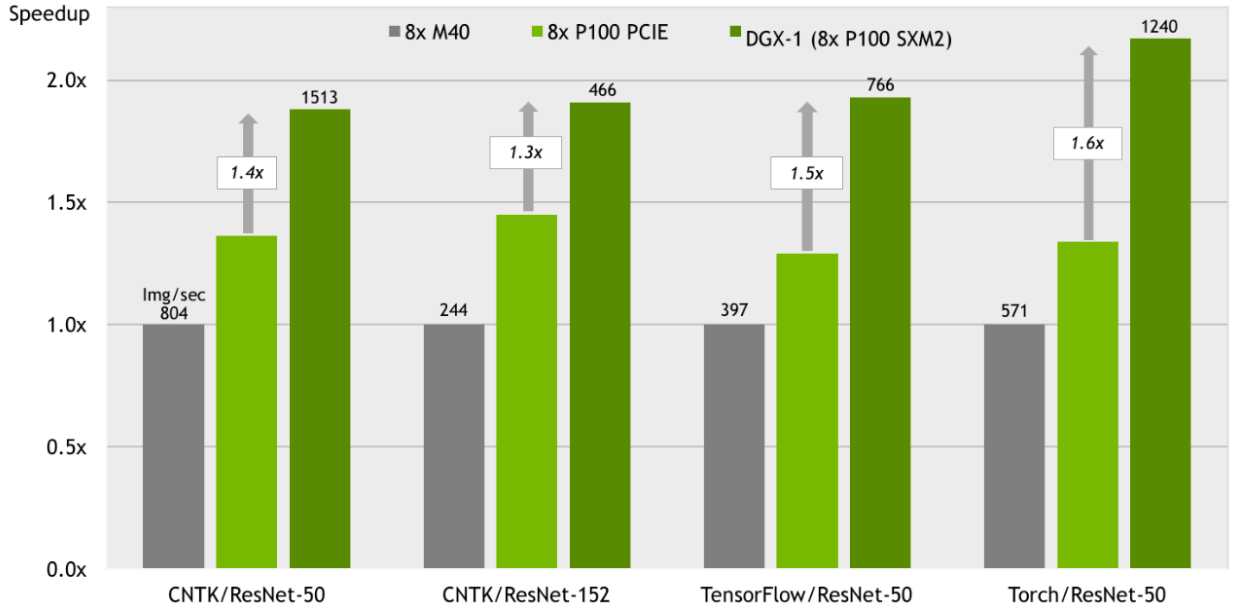

Рис. 27. Ускорение глубокого обучения

В имеющемся кластере вычислительного центра СПбГУ, на котором производились тесты GPU, имеется также программно-аппаратная архитектура параллельных вычислений **CUDA** (англ. Compute Unified Device Architecture), с помощью которой возможно самостоятельно организовывать доступ к набору инструкций графического ускорителя и управлять его памятью [61, 62].

2.4 Выбор решения, моделирующего деятельность правого полушария

TrueNorth. Все существующие алгоритмы (созданные для классической архитектуры фон-Неймана) сложно адаптируются для

использования с TrueNorth, например, чтобы запустить сверточную нейросеть требуется глубокая адаптация под нейроморфную архитектуру. В конечном итоге такая нейросеть существенно отличается от популярных вариантов: используется бинарный нейрон, а не 32-битный; тринарный синапс (-1, 0, 1). Полученный алгоритм возможно распараллелить, но данный вопрос остается неясен, т.к. все привычные средства для данной архитектуры не подходят.

Оценить производительность чипа очень сложно, т.к. типичные тесты вроде HPL, используемые для суперкомпьютеров, не адаптированы под нейроморфную архитектуру, и производительность невозможно измерить в единицах измерения FLOPS. При этом, теоретически один чип это 1 млн. процессоров, каждый из которых обрабатывает информацию с 256 каналов входа за один такт, что предполагает очень большое число. Вычисления в TrueNorth возможно измерить использованием синаптических операций в секунду (SOPS). Так, чип обеспечивает 46 миллиардов SOPS на ватт для типичной сети и 400 миллиардов SOPS на ватт для сетей с высокими скоростями спайков и большим количеством активных синапсов, тогда как сегодняшний самый производительный суперкомпьютер достигает 4,5 миллиарда FLOPS на ватт".

Эффективность чипа можно оценить с помощью работы нейронных сетей (на которые чип «заточен). Было принять прогнать TrueNorth на тесте CIFAR10 (датасет изображений) в задаче распознавания образов и сравнить с существующими алгоритмами [63]. Устройство способно классифицировать изображения со скоростью примерно 1200-2600 кадров в секунду, т.е. около 6000 FPS/Вт. Лучшие алгоритмы глубокой сверточной нейросети способны дать около 96,73% правильных ответов, что касается TrueNorth – 89,32% на восьми параллельно работающих чипах. В условиях университетского исследования имеется возможность приобрести всего один чип [64].

Использование TrueNorth ведет за собой ряд проблем: установка и подключения чипа в вычислительный кластер, увеличение времени на

кодирование задачи с учетом адаптации алгоритмов, обучение сотрудников, узкий профиль специализации (нейроморфные архитектуры), несбалансированность производительности между вычислительными компонентами установки.

Power9. Power9 имеет возможность масштабирования до 24 ядер, для одного чипа это 96 потоков, т.е. 4 потока на ядро. Исследована производительность чипа на тесте CIFAR10 – 9,6 TFLOPS, в реальных условиях производительность соответствует примерно 75% от пиковой, что соответствует 7,2 TFLOPS. Учитывая производительность смоделированной части «левого полушария» равной 7,6 TFLOPS достигается сбалансированность компонентов установки.

Технология Power имеет удобную платформу углубленного обучения IBM PowerAI для повышения доступности и производительности машинного обучения в условиях искусственного интеллекта. Платформа достаточно быстро внедряется (в виде двоичных пакетов), что является большим плюсом для адаптации алгоритмов, по сравнению с TrueNorth. PowerAI содержит среды и библиотеки машинного обучения, которые уже настроены для пиковых нагрузок. Примечательно, что платформа создана специально для поддержания нескольких архитектур аппаратного обеспечения: GPU, CPU и их взаимосвязь CPU+GPU NVLink. Данное решение подходит под наши ресурсные условия [65].

Глава 3. Тестирование производительности компонентов

3.1 Тестирование производительности сети

Одним из условий согласованности разных архитектур правого и левого полушарий является эффективная связь между ними. Каждая из частей установки и сеть должны иметь сравнимые скорости, т.к. оперируя разными скоростями, обмен сообщениями между архитектурами может быть невозможен, например, при передаче данных от одной части другой – вторая может «не услышать» сигнал передачи. Для достижения эффективной коммуникации узлов нами принято использовать высокоскоростную коммутируемую сеть для суперкомпьютеров с организацией RDMA (Remote Direct Memory Access, удалённый прямой доступ к памяти), имеющую высокую пропускную способность, малую латентность (задержку) и низкую нагрузку на процессор – InfiniBand [66].

3.1.1 InfiniBand

InfiniBand (IB) в связке с RDMA осуществляет передачу данных из памяти компьютера, находящегося удаленно, в локальную память другого компьютера, посылаемого запрос, непосредственно сетевым контроллером. Центральный процессор удаленного узла не задействован. Важно, что RDMA передает данные без лишней буферизации (метод zero-copy), а так же не нагружает узел, к памяти которого обращаются (рис. 28) [67]. Следствие – снижение латентности доступа. Благодаря RDMA значительно возрастает производительность.

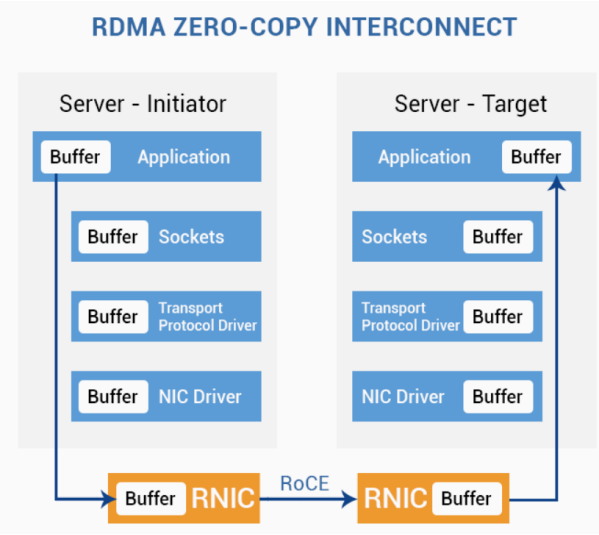

Рис. 28. Пересылка сообщений с помощью RDMA

Особенности InfiniBand:

- доступ к памяти RDMA (write/read в память получателя);
- пересылка сообщений по каналам (сообщение приходит в заранее отведенный буфер получателя);
- транзакционные операции;
- передача нескольким получателям;
- атомарная операция в память удаленного компьютера.

Архитектура InfiniBand представлена на рисунке 29.

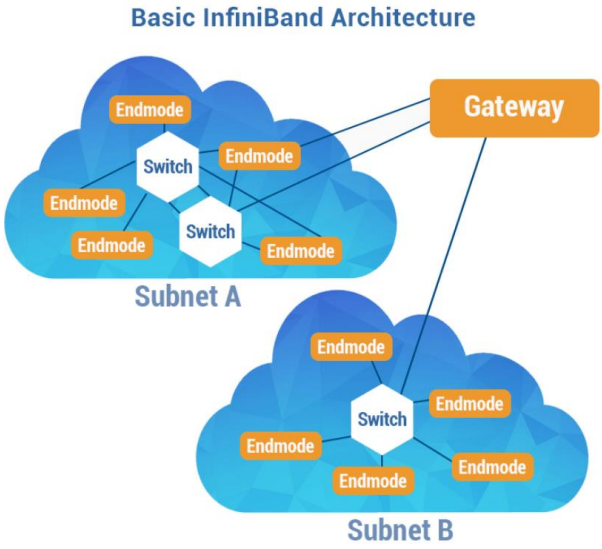

Рис. 29. Архитектура InfiniBand

Латентность и пропускная способность – самые важные параметры при сравнении коммутационных сетей. Компания Mellanox (производитель InfiniBand) предоставляет данные тестирования сетей с разной пропускной способностью и задержкой, представленные на рис. 30 [66].

	Mellanox 56 Гбит/с FDR IB	Intel 40 Гбит/с QDR IB	Intel 10 GbE NetEffect NE020
Пропускная способность	6,8 Гб/с	3,2 Гб/с	1,1 Гб/с
Задержка	0,7 мкс	1,2 мкс	7,22 мкс
Скорость передачи сообщений	137 млн сообщений в секунду	30 млн сообщений в секунду	1,1 млн сообщений в секунду

Рис. 30. Сравнение коммутационных сетей

На рисунке 30 видно, что сети с организацией InfiniBand – Mellanox 56 Гбит/с FDR IB и Intel 40 Гбит/с QDR IB превосходят 10Гиб/с сеть Ethernet, что актуально для высоконагруженных и параллельных систем. В таблице 1 представлена производительность поколений InfiniBand.

Таблица 1. Производительность InfiniBand

Поколение:	**SDR**	**DDR**	**QDR**	**FDR-10**	**FDR**	**EDR**	**HDR**
Эффективная пропускная способность, Гбит/с, на 1x шину [68]	2	4	8	10	14	25	50
Эффективные скорости для 4x и 12x шин, Гбит/с [69]	8, 24	16, 48	32, 96	41, 123	54, 163	100, 300	200, 600
Кодирование (бит)	8/10	8/10	8/10	64/66	64/66	64/66	

Типичные задержки, мкс	5	2.5	1.3	0.7	0.7	0.5	
Год появления	2001, 2003	2005	2007		2011	2014	~2017

Mellanox стал продавать оборудование с пропускной способностью 200/600 Гбит, что является полезной перспективой в применении данной технологии.

3.1.2 Intel MPI Banchmark

Для тестирования сети InfiniBand и Ethernet в реальных условиях были использованы тесты Intel MPI Banchmark. Тестирование проводилось на виртуальном кластере Вычислительного центра (ВЦ) СПбГУ.

Ресурсная база ВЦ СПбГУ

- Блейд-сервер[3] HP BL460G7:
- Система хранения HP:
  - StorageWorks P4500 G2
  - 240 TB (120 x 2TB SAS HDD)
- Процессор: Intel Xeon CPU E5-2690 v4 @ 2.60GHz
  - Sockets 2
  - Core(s) per socket 14
  - Thread(s) per core 2
- ОЗУ: 256GB RAM
- Сеть: 10GbE, QDR IB
- Два коммутатора ProCurve E6600-48G-4XG

[3] Сервер, компоненты которого вынесены и обобщены в корзину (шасси, предоставляющее доступ к общим компонентам, например, контроллер, блок питания)

    - (48 x 1GbE и 4 x 10GbE)
- ОС: Fedora 27 Linux
- Compiler Open MPI 1.6.5
- GPU NVIDIA Tesla P100 PCIe 16GB
    - Number of GPUs 2
    - CUDA 6.5

MPI (Message Passing Interface, интерфейс передачи сообщений) — стандарт пересылки сообщений в кластерах и суперкомпьютерах (преимущественно для систем с распределенной памятью). MPI предоставляет основные функции, необходимые для тестирования параллельных вычислений. Оценки эффективности используемых алгоритмов можно прочесть в работах, приведенных в списке литературы [70, 71, 72]. В тестировании использовалась версия MPI 3.1.

Intel MPI Benchmark представляет собой библиотеку с набором различных тестов на производительность кластерной системы, узлов, задержку сети и пропускную способность. Распространяется в виде проекта с открытым исходным кодом, поэтому есть возможность использовать тесты в различных архитектурах кластеров и реализации MPI [73].

Имеется возможность проводить тесты в нескольких режимах:

- standard (по умолчанию) – эталонные тесты выполняются в одной группе процессов;
- multiple – эталонные тесты выполняются в нескольких группах процессов.

А также в нескольких классификациях:

- single Transfer – одиночная передача;
- parallel Transfer – параллельная передача;
- collective benchmarks – коллективные тесты.

Более подробное описание методов можно прочесть в источнике [74].

3.1.3 Результаты тестирования

Приведены основные тесты для сравнения пропускной способности и латентности сетей InfiniBand 40 Гбит/с и Ethernet 10 Гбит/с.

Передача данных в условиях теста:

- минимальная длина сообщений: 0 байт;
- максимальная длина сообщений: 4194304 байт.

Коммуникация «точка-точка». Тест PingPong

На рисунке 31 показана схема тестирования сети методом PingPong, где MPI_Send – функция передачи сообщения, а MPI_Recv – приёма. Тест используется для измерения латентности и пропускной способности одного сообщения, отправленного между двумя процессами. Процесс1 передает сообщение процессу2, процесс2 принимает сообщение и после этого передает сообщение обратно процессу1 (непрерывный отскок сообщений друг от друга).

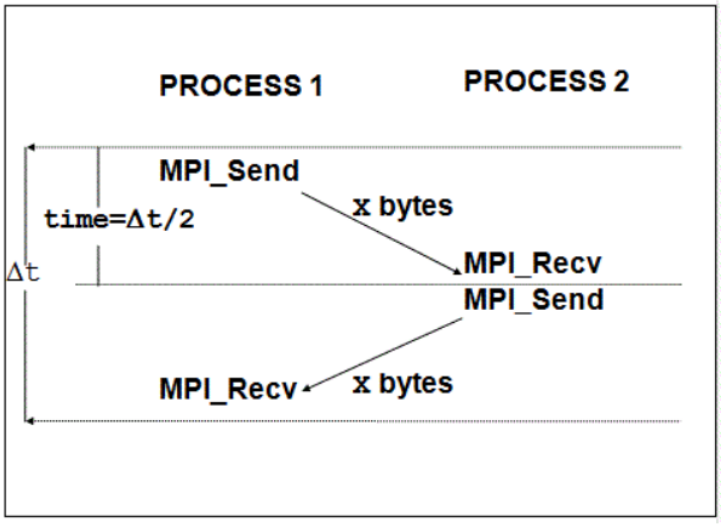

Рис. 31. PingPong pattern

На рисунках 31-35 представлены результаты тестирования сети Ethernet и InfiniBand,

где bytes – длина сообщения в байтах,

repetitions – итерации,

t – латентность в мкс,

Mbytes/sec – пропускная способность в Мб/с.

Пропускная способность – отношение времени (мкс) к длине сообщения (байт).

```
# Benchmarking PingPong
# #processes = 2
#----------------------------------------------------
       #bytes #repetitions      t[usec]   Mbytes/sec
            0         1000        25.60         0.00
            1         1000        24.30         0.04
            2         1000        24.21         0.08
            4         1000        24.14         0.17
            8         1000        24.11         0.33
           16         1000        23.79         0.67
           32         1000        23.45         1.36
           64         1000        22.98         2.78
          128         1000        22.39         5.72
          256         1000        30.63         8.36
          512         1000        41.44        12.36
         1024         1000        54.62        18.75
         2048         1000        68.30        29.98
         4096         1000        77.39        52.93
         8192         1000        81.30       100.76
        16384         1000        81.94       199.96
        32768         1000       103.06       317.94
        65536          640       214.29       305.83
       131072          320       245.97       532.88
       262144          160       401.47       652.96
       524288           80       612.60       855.84
      1048576           40      1068.06       981.76
      2097152           20      1961.44      1069.19
      4194304           10      3757.85      1116.15
```

Рис. 32. Результаты тестирования сети Ethernet 10Гбит/с

```
#---------------------------------------------------
# Benchmarking PingPong
# #processes = 2
#---------------------------------------------------
       #bytes #repetitions      t[usec]   Mbytes/sec
            0         1000         1.39         0.00
            1         1000         1.43         0.70
            2         1000         1.42         1.41
            4         1000         1.39         2.87
            8         1000         1.41         5.69
           16         1000         1.39        11.49
           32         1000         1.39        22.96
           64         1000         1.51        42.52
          128         1000         2.02        63.38
          256         1000         2.25       113.81
          512         1000         2.52       202.84
         1024         1000         2.91       351.39
         2048         1000         3.68       557.17
         4096         1000         4.55       900.68
         8192         1000         6.49      1262.17
        16384         1000         8.86      1850.06
        32768         1000        12.96      2528.87
        65536          640        21.40      3062.34
       131072          320        38.26      3425.98
       262144          160        71.52      3665.21
       524288           80       138.15      3795.19
      1048576           40       271.60      3860.77
      2097152           20       537.48      3901.79
      4194304           10      1071.40      3914.77
```

Рис. 33. Результаты тестирования InfiniBand

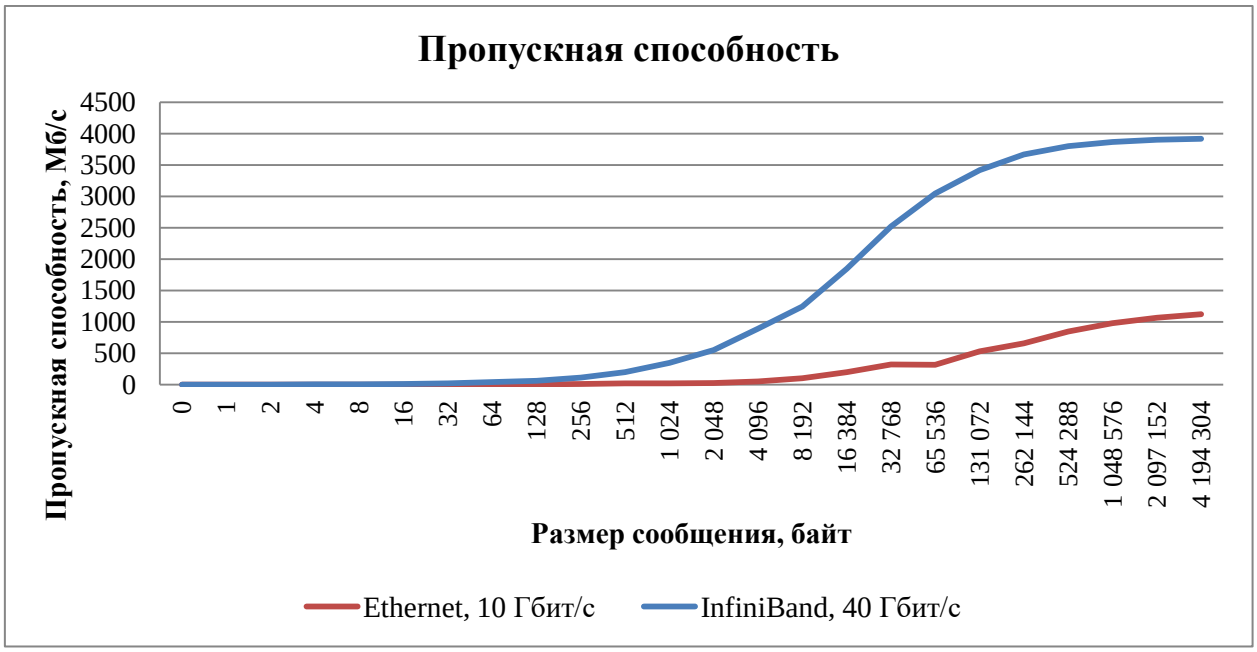

Рис. 34. Сравнение пропускной способности коммуникационных сетей на тесте PingPong

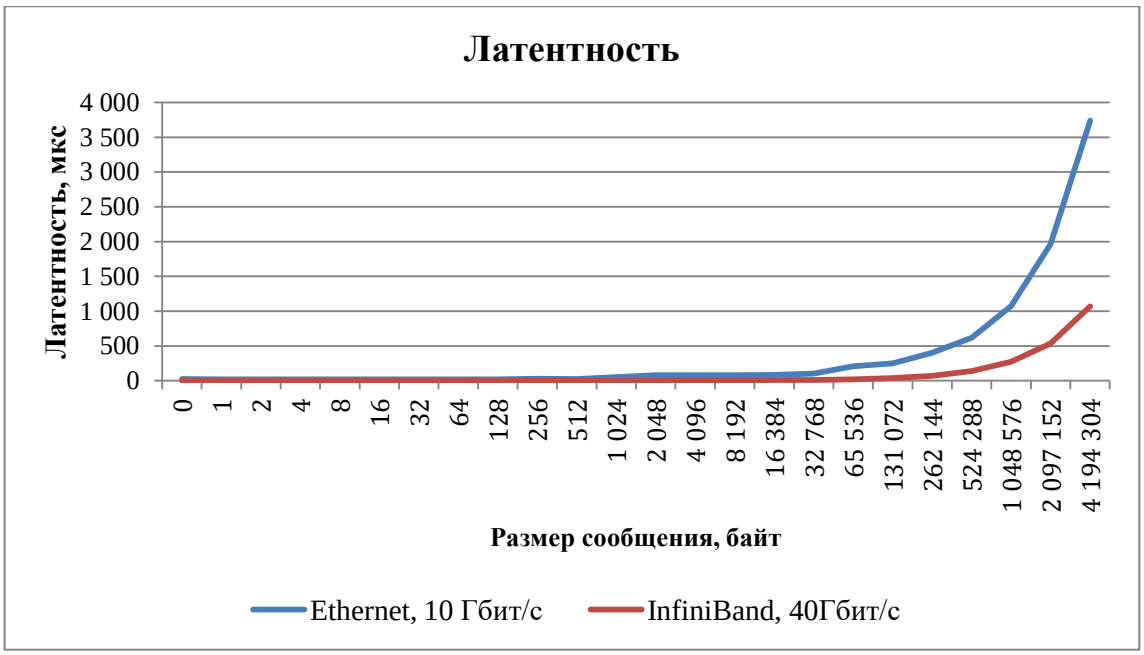

Рис. 35. Сравнение задержки коммуникационных сетей на тесте PingPong

Тесты PingPong определили, что сеть InfiniBand обладает лучшей пропускной способностью и наименьшей латентностью сети в сравнении с Ethernet.

Барьерная синхронизация. Тест Barrier

Функция Barrier блокирует вызывающий процесс до тех пор, пока все процессы в коммуникаторе не вызовут его; то есть вызов возвращается в любом процессе только после того, как все члены коммуникатора вошли в вызов. Выполнение следующей за Barrier инструкции каждая задача начнет одновременно. На рисунках 36-37 представлены результаты тестирования сетей Ethernet 10 Гбит/с и InfiniBand 40 Гбит/с методом Barrier.

```
#---------------------------------------------------
# Benchmarking Barrier
# #processes = 2
#---------------------------------------------------
 #repetitions  t_min[usec]  t_max[usec]  t_avg[usec]
         1000        24.55        24.55        24.55
```

Рис. 36. Результат теста Barrier для сети Ethernet 10Gbit/s

```
#---------------------------------------------------
# Benchmarking Barrier
# #processes = 2
#---------------------------------------------------
 #repetitions  t_min[usec]  t_max[usec]  t_avg[usec]
         1000         1.21         1.21         1.21
```

Рис. 37. Результат теста Barrier для сети InfiniBand QDR

Сеть InfiniBand показала результат лучше, чем Ethernet: задержка ниже в 20,3 раз, что является существенной разницей в условиях параллельных вычислений и многозадачности.

Исходя из полученных результатов, можно сделать вывод, что сеть InfiniBand обладает лучшей пропускной способностью и наименьшей латентностью сети в сравнении с Ethernet.

3.2 Тестирование NVIDEA GPU Tesla P100

Тестирование установки с GPU проводилось с помощью бэнчмарка High Performance LINPACK.

LINPACK — библиотека, написанная на языке Фортран и адаптированная под язык C, для решения больших систем линейных алгебраических уравнений методом LU-разложения[4]. Кроме того, результаты теста используются в составлении списка Top500 (рейтинг мощнейших вычислительных систем) [76]. Производительность определяется количеством "полезных" вычислительных операций над числами с плавающей точкой в секунду, выражается в GFLOPS/сек [75].

[4] LU-разложение — это представление матрицы A в виде L*U, где L — нижнетреугольная единичная матрица, а U — верхнетреугольная матрица, используется в решении систем алгебраических уравнений, вычисления определителя, обратной матрицы и др. LU-разложение – модификация метода Гаусса.

HPL (High Performance LINPACK) – тест на производительность кластера, реализация на языке Си (решение СЛАУ с двойной точностью (64 бита) для компьютеров с распределенной памятью. HPL требует наличия реализации интерфейса MPI, а вычисления на каждом процессоре - с помощью процедур BLAS (*Basic Linear Algebra Subprograms* выполняет основные операции линейной алгебры, такие как умножение векторов и матриц), например, используя библиотеку Intel MKL.

Тестирование производилось на ноде вычислительного кластера, имеющего 2 карты GPU NVIDIA GP100GL [Tesla P100 PCIe 16Gb] и CPU, имеющего характеристики, представленные на рисунке 38.

```
[sinel@w3 ~]$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                56
On-line CPU(s) list:   0-55
Thread(s) per core:    2
Core(s) per socket:    14
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 79
Model name:            Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz
Stepping:              1
CPU MHz:               2594.016
CPU max MHz:           3500,0000
CPU min MHz:           1200,0000
BogoMIPS:              5188.03
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              256K
L3 cache:              35840K
NUMA node0 CPU(s):     0-13,28-41
NUMA node1 CPU(s):     14-27,42-55
```

Рис.38. Конфигурация CPU

Для получения максимальных вычислительных характеристик кластера необходимо выбрать соответствующие значения для ряда параметров. Настраиваемые параметры попадают в две основные категории: параметры задачи, описанные в файле HPL.dat, и параметры среды. Перед каждым запуск теста HPL, необходимо убедиться, что кластер не запускает сторонних высоконагруженных приложений, которые могут повлиять на результаты эталонного теста. В случае использования GPU команда nvidia-smi

вызывается перед каждым тестовым прогоном для контроля занимаемой памяти (рис. 39).

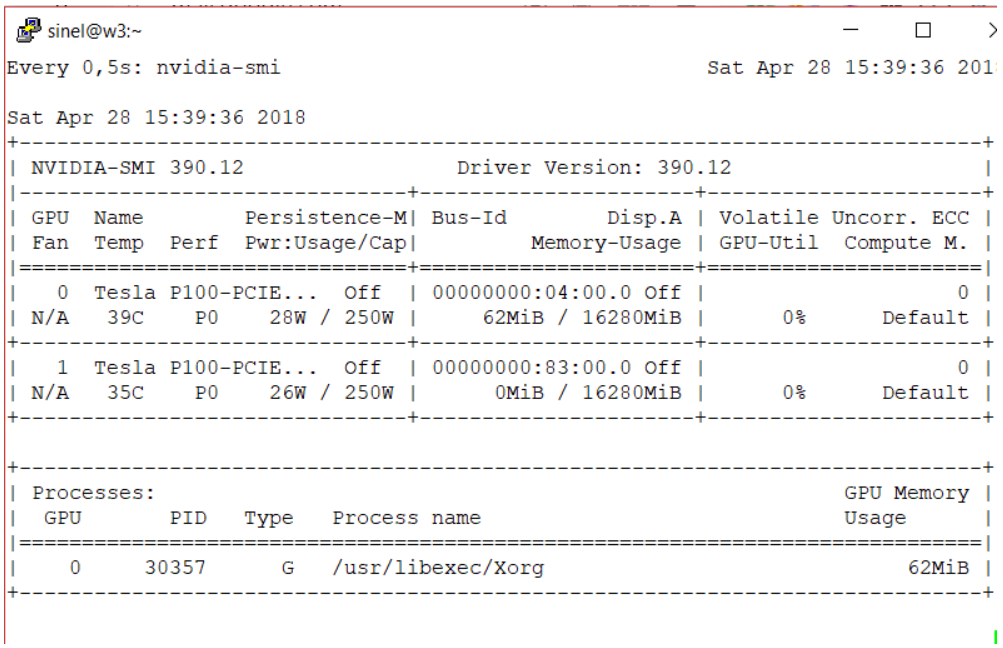

Рис. 39. Команда watch -n 0,5 nvidia-smi

Для контроля загруженности ядер используется команда htop (рис. 40).

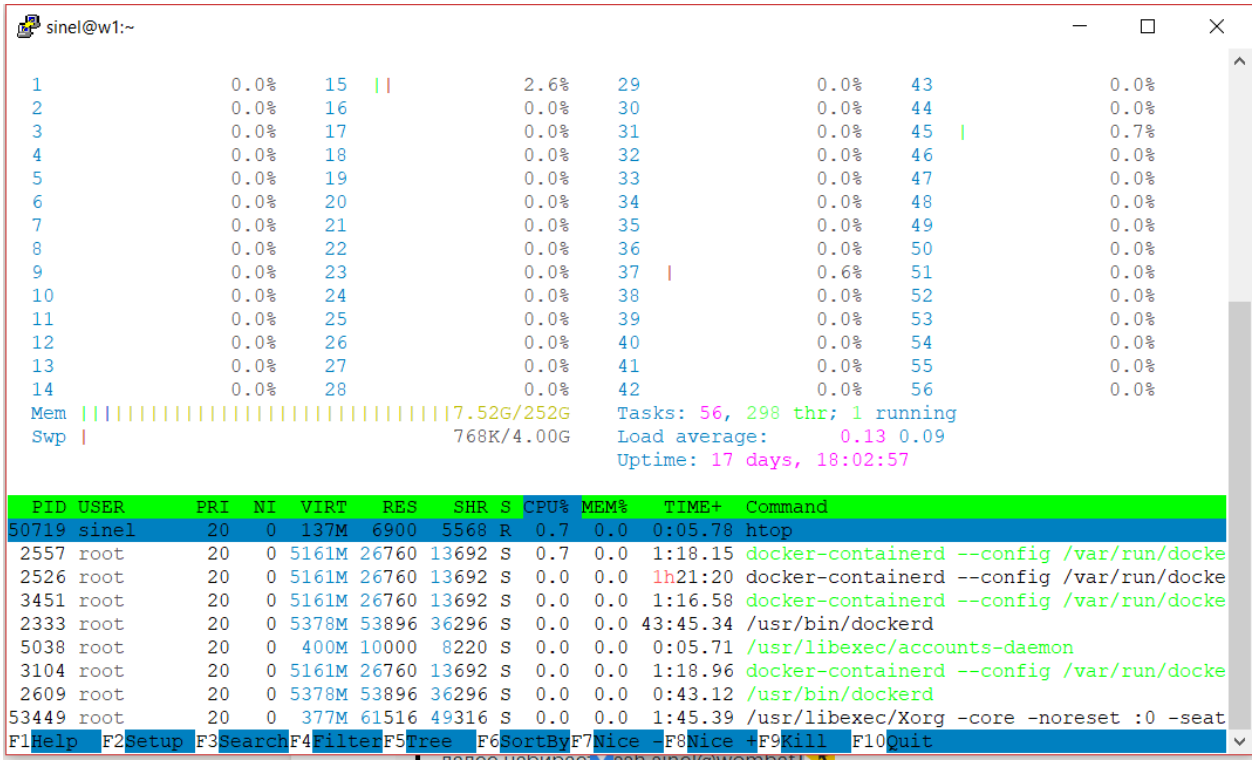

Рис. 40. Команда htop. Загруженность ядер

3.2.1 Настройка параметров задачи

Ns. Наиболее важным параметром при настройке HPL является размер решаемой задачи. Размер матрицы определяет загрузку вычислителей, чем больше матрица, тем больше загружена память. Оптимально, когда память вычислителя загружена более чем на 90%. Стоит отметить, что даже в случае выполнения HPL на графическом процессоре, размер задачи ограничен размером оперативной памяти кластера, а не памяти GPU. Поэтому в подборе параметра Ns поможет сам бенчмарк, который при запуске выводит информацию о занимаемом объеме памяти, например: *Per-Process Host Memory Estimate: 16.16 GB (MAX) 16.16 GB (MIN).*

Для данной конфигурации кластера максимальная производительность достигается, когда размер матрицы Ns на 20% больше, чем доступная память GPU (Ns = 68352). Но для окончательного запуска выбирается значение, позволяющее использовать ≈ 96% доступной памяти графического процессора, что является оптимальным решением: Ns = 62976. Результаты производительности, в зависимости от различного размера матрицы представлены на рисунке 41.

Рис. 41. Производительность в зависимости от Ns

NBs. Размер блоков, на которые делится матрица решаемой задачи, имеет значительное влияние на производительность кластера. Исследования показали, что при использовании GPU следует указывать большее значение параметра, чем при вычислении на CPU. Максимальные результаты достигаются при использовании значений для NB, кратного 64 (допустимо 32). Кроме того, параметр зависит от Ns, если размер матрицы Ns делится на размер блока без остатка, достигаются оптимальные результаты.

Для данной задачи NBs = 384, полученное значение удовлетворяет выражениям: 384 mod 32 = 0, 384 mod 64 = 0, 62976 mod 384 = 0. Результаты представлены на рисунке 42.

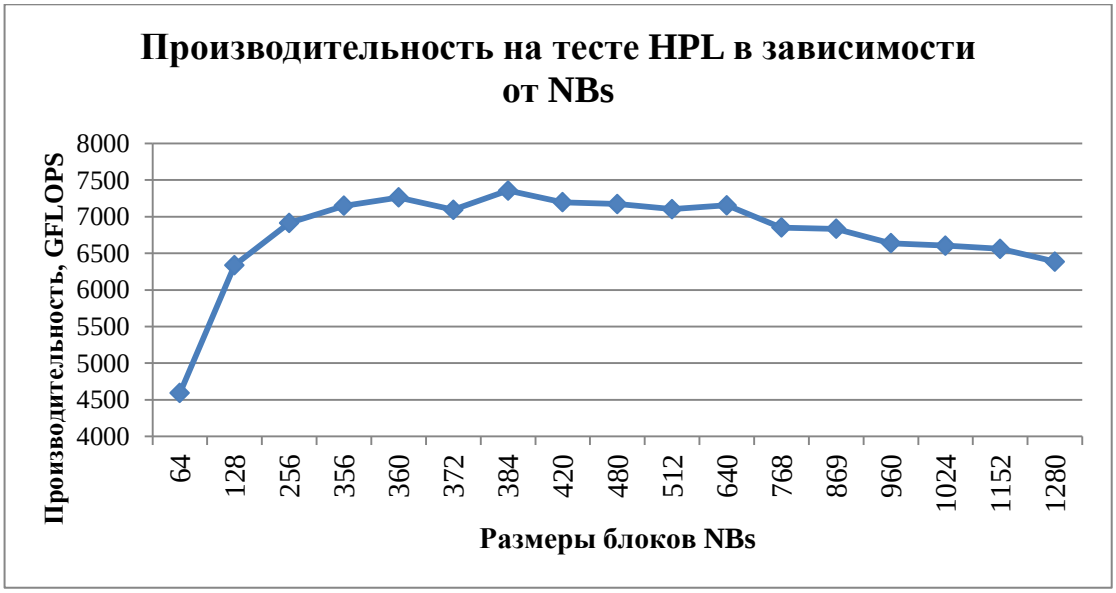

Рис. 42. Производительность в зависимости от размера блоков

P, Q. Количество строк и столбцов каждой сетки на которой считается бенчмарк. Главное правило при настройке: P*Q = число MPI процессов (и = числу GPU, в нашем случае). В нашем случае возможны только два варианта, которые отвечают этому требованию: P = 1 и Q = 2; P = 2 и Q = 1. В этом случае использование P = 1 возможно только в том случае, когда размер

матрицы, который является размером задачи, не превышает память графического процессора, т.е. оптимальная конфигурация P = 2 и Q = 1.

Конфигурации для файла, содержащего настройки параметров, которые соответствуют лучшему результату, представлены на рисунке 43.

```
6            device out (6=stdout,7=stderr,file)
1            # of problems sizes (N)
62976        Ns
1            # of NBs
384          NBs
1            PMAP process mapping (0=Row-,1=Column-major)
1            # of process grids (P x Q)
2        Ps
1        Qs
16.0         threshold
1            # of panel fact
2        PFACTs (0=left, 1=Crout, 2=Right)
1            # of recursive stopping criterium
4          NBMINs (>= 1)
1            # of panels in recursion
2            NDIVs
1            # of recursive panel fact.
0 1 2        RFACTs (0=left, 1=Crout, 2=Right)
1            # of broadcast
2 0 2          BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1            # of lookahead depth
0            DEPTHs (>=0)
1            SWAP (0=bin-exch,1=long,2=mix)
192          swapping threshold
1            L1 in (0=transposed,1=no-transposed) form
0            U  in (0=transposed,1=no-transposed) form
1            Equilibration (0=no,1=yes)
8            memory alignment in double (> 0)
```

Рис. 43. Конфигурация файла с настройками параметров

3.2.2 Настройка параметров среды

GPU_DGEMM_SPLIT. Параметр, который определяет какая часть задачи будет выполняться на GPU, а какая на CPU. Наибольшая производительность достигается при значении параметра > 0,9. Значение = 1.00 означает, что 100% вычислений выполнено с помощью графических процессоров. Когда HPL запущен на P100 мы устанавливаем GPU DGEMM

SPLIT = 1.00, что означает, что CPU обрабатывают служебные данные и перемещают данные для поддержки мощности графических процессоров.

CPU_CORES_PER_RANK. Количество физических ядер в узле, разделенных рядами на узел (RANK PER NODE = 2). Значение параметра равное 28 позволяет полностью загрузить CPU, однако, поскольку DGEMM SPLIT = 1.00, этот параметр не показывает максимальной эффективности. Оптимальная производительность достигается при использовании процессора CORES PER RANK = 20.

Power limit. Параметр GPU, определяющий максимальное энергопотребление GPU. В случае P100 ограничение мощности должно быть между 150 и 250 Вт. Фактический контроль за потребленной энергией осуществляется путем динамического изменения **application clocks**. Application clocks состоит из двух компонент Graphics и Memory, которые определяют скорость GPU при запуске приложений на графическом процессоре. Для графического процессора P100 имеется только одна опция для application memory, а значение graphics clocks может быть изменено. Доступные значения для этого параметра, такие как ограничение мощности, определяются для каждого конкретного GPU. Изменение тактовых импульсов приложений полезно при получении максимального значения коэффициента производительности/энергопотребления, значение которого является приоритетом при построении системы HPC. Следует отметить, что для теста HPL не рекомендуется использовать максимальное значение для application clocks.

Команда для изменения power limit: sudo nvidia-smi -pl X.

Команда для изменения application clocks: sudo nvidia-smi -ac Memory,Graphics.

Для данной задачи выбраны настройки nvidia-smi -ac 715,1290 и nvidia-smi -pl 250.

3.2.3 Результаты тестирования производительности установки с GPU NVIDIA Tesla P100

Несмотря на то, что все вычисления выполняются на графическом процессоре, теоретическая пиковая производительность системы основана на формуле:

$P = CPU_{max} + GPU_{max}$,

где CPU_{max} и GPU_{max} обозначают максимальную производительность.

Процессоры $CPU_{max} = 2.60 \times 2 \times 14 \times 2 \times 16 \approx 2329,6$ GFLOPS

Графические процессоры $GPU_{max} \approx 2 \times 4812,8$ GFLOPS $\approx 9625,6$ GFLOPS

P = 2329, 6 GFLOPS + 9625, 6 GFLOPS = 11955, 2 GFLOPS.

Окончательный результат тестирования производительности с помощью теста HPL, включая настройку параметров и оптимизацию окружения, составляет **7609 GFLOPS**, т.е. 63,6% от максимальной производительности.

Глава 4. Конфигурация экспериментальной установки, моделирующей деятельность мозга

Минимальная реализуемая конфигурация установки

- GPU: 2x Tesla P100
- память GPU: 16 GB на GPU
- CUDA 6.5
- Чип IBM Power9
- CPU: Dual 20-core Intel Xeon CPU E5-2690 v4 @ 2.60GHz
- ОЗУ: 256 GB
- Система хранения данных:
- HP StorageWorks P4500 G2
- 240 TB (120 x 2TB SAS HDD)
- Сеть передачи данных: IB QDR 4X 40Gbit/sec
- PCI express 3.0
- ОС: Fedora 27 Linux
- максимальные требования к энергии: 3200W
- операционная температура: 10 – 35 oC

В перспективе рассматривается использование DGX-1 с 8 картами Tesla P100 или Tesla V100, InfiniBand EDR 100Gbit/s, NVlink.

Данную установку можно использовать в различных задачах:

- Медицина: анализ истории болезней пациента, диагностирование диагноза, подбор индиидуального лечения.
- Безопасность жизнедеятельности: предсказание перебоев с электричеством, угроз цунами, землетрясений, эффективная организация эвакуации, предсказывание автомобильного движения по городу.

- Образование. Отслеживание неуспеваемости школьников, анализ причин, предложения по индивидуальной программе обучения.
- Кулинария. Создание рецептов и меню индивидуально для каждого. Оценка качества продуктов.
- Бизнес. Прогноз успешных сделок, учет рисков и др.

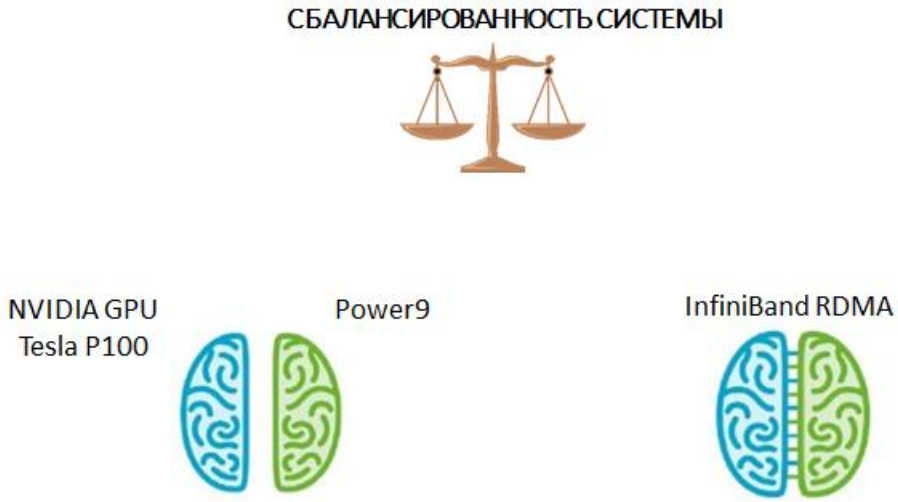

Рис. 44. Модель установки

Выводы

В ходе проделанной работы были получены следующие результаты:

1) выполнен анализ существующих решений среди проектов по моделированию мозга, выявлены их особенности и недостатки в задачах моделирования мыслительных процессов, большинство проектов направлены на исследование мозга с точки зрения анатомического, биологического строения. Описанные проекты имеют высокую финансовую поддержку для обеспечения лабораторий.

2) выполнен анализ архитектур TrueNorth и Power9, используемых для моделирования деятельности мозга в задачах искусственного интеллекта. Для дальнейшего использования в конфигурацию экспериментальной установки выбран чип Power9, который позволит моделировать деятельность «правого полушария».

3) исследованы особенности функциональной асимметрии головного мозга человека, доказывающие анатомические и функциональные различия разных полушарий, что объясняет решение создания экспериментальной установки с учетом описанных особенностей;

4) проведено тестирование компонентов экспериментальной установки: GPU NVIDIA Tesla P100, отвечающего за аналитическое полушарие, а также сетей Ethernet и InfiniBand (решено осуществлять связь между компонентами по сети InfiniBand QDR);

5) предложена оптимальная минимальная конфигурация экспериментальной установки для задач моделирования деятельности мозга с применением альтернативного подхода, учитывающего асимметрию головного мозга.

Заключение

Исследования в области мозга приобретают глобальные масштабы, вышеописанные проекты пытаются ответить на вопросы, которые во многом определяют нас: как работает наш мозг, сознание, в чем причины психических расстройств и нейродегенеративных заболеваний. Сейчас подобные исследования – крупные наднациональные проекты, но для их успешной реализации необходима эффективная инфраструктура (банки образцов, масштабные производства средств для научных исследований и т. п.), огромная вычислительная мощность, которую исследовательские группы наращивают до допустимого уровня.

Решение, описанное в данной работе, предполагает комбинацию нескольких архитектур, учитывающих функциональную асимметрию головного мозга: «правое полушарие», основанное на деятельности нейронных сетей, на базе специально созданного для задач искусственного интеллекта процессора Power9 и «левое полушарие» с мощной аналитической составляющей из GPU NVIDIA Tesla P100. Данная конструкция является сбалансированной по производительности, а также скорости передачи данных благодаря сети InfiniBand QDR RDMA, что является крайне важным в условиях моделирования головного мозга, ибо дисбаланс в человеческом мозге – это и есть деменция (психическое/когнитивное заболевание), поэтому важно не выстроить «машину шизофреника». Подход, моделирующий мышление, выбран по причине нарастающей проблемы для решения большого количества задач, где быстродействия стандартных алгоритмов не хватает.

Отличительной особенностью данного проекта СПбГУ является сотрудничество с Национальным медицинским исследовательским центром психиатрии и неврологии имени В.М. Бехтерева, на данных которого планируется обучать систему (данные МРТ у пациентов в норме, патологии и др., которые необходимо классифицировать, кластеризовать и привести к

обучающим случаям). Помимо медицинской сферы (диагностика когнитивных заболеваний на ранних стадиях) систему, построенную по принципам человеческого мышления, можно будет использовать в совершенно разных областях: финансы, промышленность, анализ данных, маркетинг, энергетика и др. Кроме того, многие сферы коммерческой деятельности сталкиваются с проблемой решения задач в режиме реального времени, в данном случае важным моментом является высокая производительность системы, которую возможно нарастить по мере усложнения решаемых задач. Подход на основе комбинации аналоговых и цифровых систем, наподобие человеческого мозга, выглядит перспективным.

Список литературы

1. Психические расстройства и заболевания [Электронный ресурс] URL: http://psyhodisease.ru/sindrom-kapgra-ili-bred-otricatelnogo-dvojnika.html (дата обращения: 18.10.2017)
2. Митио Каку. Будущее разума. М.: Альпина Нон-фикшн, 2015. 502 с.
3. Human brain Project. Brain Simulation [Электронный ресурс] URL: https://www.humanbrainproject.eu/en/brain-simulation (дата обращения: 18.10.2017)
4. Artificial brains [Электронный ресурс] URL: http://www.artificialbrains.com (дата обращения: 19.10.2017).
5. Blue Brain Project. [Электронный ресурс] URL: https://bluebrain.epfl.ch/page-56882-en.html (дата обращения: 19.10.2017)
6. National Geographic [Электронный ресурс] URL: http://www.nat-geo.ru/science/813167-sozdana-samaya-slozhnaya-kompyuternaya-model-golovnogo-mozga (дата обращения: 21.10.2017)
7. Reconstruction and Simulation of Neocortical Microcircuitry [Электронный ресурс] URL: https://goo.gl/79FVBM (дата обращения: 30.10.2017)
8. Club of Amsterdam Journal [Электронный ресурс] URL: http://www.clubofamsterdam.com/cat_content.asp?contentid=844&catid=138#article02 (дата обращения 2.11.2017)
9. Human Brain Project [Электронный ресурс] URL: https://www.humanbrainproject.eu/ (дата обращения: 18.10.2017)
10. Brain Simulation Platform [Электронный ресурс] URL: https://www.humanbrainproject.eu/en/brain-simulation/brain-simulation-platform/ (дата обращения: 2.11.2017)
11. Allen Human Brain Atlas [Электронный ресурс] URL: http://www.brain-map.org/ (дата обращения: 13.11.2017)

12. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture [Электронный ресурс] URL: https://www.researchgate.net/figure/303558797_fig3_Figure2-Parcellation-scheme-of-the-human-brain-in-the-Brainnetome-Atlas-The-MPM-for (дата обращения 30.11.2017)

13. Geektimes. Структуру мозга, различимую при сканировании, связали с качествами личности человека [Электронный ресурс] URL: https://geektimes.ru/post/263816/ (дата обращения 15.11.2017)

14. Brenner, S. The Genetics of Caenorhabditis elegans. Genetics 77, 1974. p. 71–94.

15. The Symphony Inside Your Brain [Электронный ресурс] URL: https://directorsblog.nih.gov/2012/11/05/the-symphony-inside-your-brain/ (дата обращения: 10.10.2017)

16. Компания IBM. Когнитивные вычисления – работа быстрее мысли [Электронный ресурс] URL: https://habrahabr.ru/company/ibm/blog/276855/ (дата обращения 10.05.2017). .

17. Создание системы ответов на вопросы на естественном языке с помощью служб IBM Watson и Bluemix [Электронный ресурс] URL: https://www.ibm.com/developerworks/ru/library/cl-watson-films-bluemix-app/ (дата обращения 13.12.2017).

18. Научная Россия. Сто профессий компьютера Ватсона. [Электронный ресурс] URL: https://scientificrussia.ru/articles/sto-professij-kompiutera-watsona (дата обращения: 13.12.2017)

19. Блог компании IBM. Когнитивная система IBM Watson: принципы работы с естественным языком [Электронный ресурс] URL: https://habrahabr.ru/company/ibm/blog/266015/ .

20. Блог компании IBM. [Электронный ресурс] URL: https://habrahabr.ru/company/ibm/blog/332070/ (дата обращения 14.12.2017).

21. IBM Watson makes treatment plan for brain [Электронный ресурс] URL: https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/ibm-watson-

makes-treatment-plan-for-brain-cancer-patient-in-10-minutes-doctors-take-160-hours (дата обращения: 25.11.2017) .

22. SiliconAngle. Watson correctly diagnoses woman after doctors were stumped [Электронный ресурс] URL: https://siliconangle.com/blog/2016/08/05/watson-correctly-diagnoses-woman-after-doctors-were-stumped/ (дата обращения: 13.12.2017).

23. Баррат Дж. Последнее изобретение человечества: Искусственный интеллект и конец эры Homo sapiens М. : Альпина нон-фикшн, 2015. 304 с.

24. Кринко О.Е. Взаимодействие доминантного и субдоминантного полушарий при выполнении простой зрительно-моторной реакции // Материалы VIII Международной студенческой электронной научной конференции «Студенческий научный форум» URL: http://www.scienceforum.ru/2016/1885/25501 (дата обращения: 20.01.2018)

25. Брагина Н. Н. Функциональные асимметрии человека / Н. Н. Брагина, Т. А. Доброхотова. — 2-е изд. перераб. и доп. — М. : Медицина, 1988. 237 с.

26. Реброва Н. П. Межполушарная асимметрия мозга человека и психические процессы / Н. П. Реброва, М. П. Чернышева. — СПб., 2004. — 96 с.

27. Мещеряков Б. Большой психологический словарь. СПб.: прайм-ЕВРОЗНАК, 2004. 672 с.

28. Москвина Н.В. Межполушарные асимметрии и индивидуальные различия человека, 2011. 480 с.

29. Nobel Prizes and Laureates [Электронный ресурс] URL: https://www.nobelprize.org/nobel_prizes/medicine/laureates/1981/sperry-lecture_en.html (дата обращения: 22.01.2018).

30. Клейнман П. Психология. Люди, концепции, эксперименты. М: МИФ, 2016. 272 с.

31. Geschwind N., Galaburda A. M. Cerebral Lateralization: biological mechanisms, associations and pathology. — Cambridge, MA : MIT press, 1987 .

32. Geschwind N. Specialization of the Human Brain // Scientific American. — 1979. — Vol. 241, no. 3. P. 180-199.

33. Мышление и сознание [Электронный ресурс] URL: http://galactic.org.ua/Xomo/m997.htm (дата обращения 25.01.2018)

34. Лауреаты Нобелевской премии: Энциклопедия: Пер. с англ.– М.: Прогресс, 1992. © The H.W. Wilson Company, 1987. [Электронный ресурс] URL: http://n-t.ru/nl/mf/sperry.htm (дата обращения 25.01.2018)

35. Антропова Л. К. Влияние межполушарной асимметрии на особенности проявления копинг-стратегий индивида в стрессовой ситуации / Л. К. Антропова, О. О. Андронникова, В. Ю. Куликов, Л. А. Козлова // Современные направления в исследовании функциональной межполушарной асимметрии и пластичности мозга. М.: Научный мир, 2010. С. 75–78. .

36. Ларина О.В., Гитун Т.В., Лауреаты Нобелевской премии, «Дом Славянской книги», 2006 г., с. 430-431.

37. Энциклопедия. Центральная нервная система [Электронный ресурс] URL: http://www.grandars.ru/college/medicina/most-mozga.html (дата обращения 29.01.2018)

38. Сигел Д. Майндсайт. Новая наука личной трансформации. М.: МИФ. 2015. 336 с.

39. Амен Д. Дж. Мозг: от хорошего к превосходному. М.: Эксмо, 2013. 400 с.

40. Анатомия и физиология [Электронный ресурс] URL: http://reabilitaciya.org/anatomiya-fiziologiya/normalnaya.html?start=5 (дата обращения: 30.01.2018)

41. Леутин Е. Н. Психологические механизмы адаптации и функциональная асимметрия мозга / Е. Н. Леутин, Е. И. Николаева. — Новосибирск : Наука, СО, 1988. 192 с. .

42. Хасанова Г.Б. Антропология: учебное пособие. М.: Кнорус, 2013. 232 с.

43. Антропова Л.К. Функциональная асимметрия мозга и индивидуальные психофизиологические особенности человека /Л.К. Антропова, О.О. Андронникова, В.Ю. Куликов и др.// Медицина и образование в Сибири. Выпуск № 3, 2011. [Электронный ресурс] URL: http://cyberleninka.ru/article/n/funktsionalnaya-asimmetriya-mozga-i-individualnye-psihofiziologicheskie-osobennosti-cheloveka (дата обращения: 2.02.2018)

44. Сигел Д. Внимательный мозг. М.:МИФ, 2015. 336 с. .

45. Эмоции и принятие решений: как чувства влияют на действия [Электронный ресурс] URL: https://scr.hse.ru/news/96133388.html (дата обращения 2.02.2018)

46. Hirshleifer D. Good Day Sunshine: Stock Returns and theWeather. Journal of Finance. 2003. 1009-1032 p.

47. Аршавский. В. В. Межполушарная асимметрия в системе поисковой активности. Владивосток: ДВО АН СССР, 1988. 136 с.

48. Ротенберг, В. С. Межполушарная асимметрия мозга и проблема интеграции культур/В. С. Ротенберг, В. В. Аршавский//Вопросы философии, 1984. С. 78-86.

49. Кукушкин В. С., Столяренко Л. Д. Этнопедагогика и этнопсихология. – Ростов-на-Дону, 2000. С. 220–224.

50. Кочетков В. В. Психология межкультурных различий: Учеб. пособие для вузов. М., 2002. С. 33.

51. Вальцев С.В. Миссия России. Национальная доктрина. М.: Книжный мир, 2011. 352 с.

52. Популярная механника. Компьютер с архитектурой человеческого мозга [Электронный ресурс] URL: https://www.popmech.ru/technologies/53343-po-obrazu-i-podobiyu-mamontov/ (дата обращения 20.01.2018)

53. Принципы фон Неймана [Электронный ресурс] URL: https://inf1.info/machineneumann (дата обращения 16.02.2018)

54. Introducing a Brain-inspired Computer TrueNorth's neurons to revolutionize system architecture [Электронный ресурс] URL: http://www.research.ibm.com/articles/brain-chip.shtml (дата обращения 18.02.2018)

55. IBM creates Corelet programming language to make software that operates like the human brain [Электронный ресурс] URL: https://www.extremetech.com/extreme/163448-ibm-creates-corelet-programming-language-to-make-software-that-operates-like-the-human-brain (дата обращения 18.02.2018)

56. A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface [Электронный ресурс] URL: http://paulmerolla.com/merolla_main_som.pdf (дата обращения 19.02.2018) .

57. Revealed: A Scale-Out Synaptic Supercomputer (NS1e-16) [Электронный ресурс] URL: https://modha.org/2015/12/revealed-a-scale-out-synaptic-supercomputer-ns1e-16/ (дата обращения: 19.02.2018)

58. Открытые системы. Искусственный мозг IBM стал крупнее [Электронный ресурс] URL: https://www.osp.ru/news/articles/2016/14/13048924/ (дата обращения: 20.02.2018).

59. Sadasivam S. K. et al. IBM POWER9 Processor Architecture // IEEE Micro. — 2017. Vol. 37, N. 2. — P. 40–51 :

60. Power9 — процессоры для больших данных [Электронный ресурс] URL: https://www.osp.ru/os/2017/03/13052698/ (дата обращения: 20.04.2018)

61. NVIDIA DGX-1 [Электронный ресурс] URL: https://www.nvidia.com/en-us/data-center/dgx-1/ (дата обращения: 3.03.2018)

62. NVIDIA DGX-1: The Fastest Deep Learning System [Электронный ресурс] URL: https://devblogs.nvidia.com/dgx-1-fastest-deep-learning-system/ (дата обращения: 3.03.2018)

63. CIFAR dataset [Электронный ресурс] URL: https://www.cs.toronto.edu/~kriz/cifar.html (дата обращения 8.03.2018)

64. Свёрточные нейронные сети работают на нейроморфных чипах [Электронный ресурс] URL: https://nplus1.ru/news/2016/09/21/truenorthcnn (дата обращения 20.03.2018)

65. IBM PowerAI [Электронный ресурс] URL: https://www.ibm.com/ru-ru/marketplace/deep-learning-platform (дата обращения 23.03.2018).

66. Mellanox technologies. Производительность InfiniBand [Электронный ресурс] URL: http://ru.mellanox.com/page/performance_InfiniBand (дата обращения 24.03.2018).

67. Эффективный перенос данных с помощью zero copy [Электронный ресурс] URL: https://www.ibm.com/developerworks/ru/library/j-zerocopy/ (дата обращения 24.03.2018).

68. InfiniBand Roadmap: IBTA - InfiniBand Trade Association [Электронный ресурс] URL: http://www.InfiniBandta.org/content/pages.php?pg=technology_overview (дата обращения: 24.03.2018)

69. Mellanox workshop [Электронный ресурс] URL: http://www.hpcadvisorycouncil.com/events/2014/swiss-workshop/presos/Day_1/1_Mellanox.pdf (дата обращения: 24.03.2018)

70. R. Thakur, W. Gropp. Improving the Performance of Mpi Collective Communication on Switched Networks, 2003

71. R. Thakur, R. Rabenseifner, W. Gropp. Optimization of Collective Communication Operations in MPICH. Int'l Journal of High Performance Computing Applications,-- 2005 — Vol 19(1) — pp. 49-66.

72. J. Pjesivac-Grbovic, T. Angskun, G. Bosilca, G. E. Fagg, E. Gabriel, J. J. Dongarra. Performance analysis of MPI collective operations. Cluster Computing — 2007 — Vol. 10 — p.127.

73. B. S. Parsons. Accelerating MPI collective communications through hierarchical algorithms with flexible inter-node communication and imbalance awareness. Ph. D. Thesis on Computer science, Purdue University, USA, 2015

74. Running Benchmarks in Multiple Mode [Электронный ресурс] URL: https://software.intel.com/en-us/imb-user-guide (дата обращения: 25.03.2018)

75. Dongarra J. J., Bunch J. R., Moler G. B., Stewart G. W. LINPACK Users' Guide. — Society for Industrial and Applied Mathematics, 1979—1993, p. 367

76. Список TOP500 самых производительных суперкомпьютеров мира [Электронный ресурс] URL: https://www.top500.org/ (дата обращения 26.03.2018)